Combiners for Backdoored Random Oracles

Balthazar Bauer[2,1], Pooya Farshim[1,2], and Sogol Mazaheri[3]

[1] DI/ENS, CNRS, PSL University, Paris, France
[2] Inria, Paris, France
balthazar.bauer@ens.fr, pooya.farshim@gmail.com
[3] Cryptoplexity, Technische Universität Darmstadt, Germany
sogol.mazaheri@cryptoplexity.de

Abstract. We formulate and study the security of cryptographic hash functions in the *backdoored random-oracle* (BRO) model, whereby a big brother designs a "good" hash function, but can also see *arbitrary functions* of its table via backdoor capabilities. This model captures intentional (and unintentional) weaknesses due to the existence of collision-finding or inversion algorithms, but goes well beyond them by allowing, for example, to search for structured preimages. The latter can easily break constructions that are secure under random inversions.

BROs make the task of bootstrapping cryptographic hardness somewhat challenging. Indeed, with only a single arbitrarily backdoored function no hardness can be bootstrapped as any construction can be inverted. However, when two (or more) independent hash functions are available, hardness emerges *even with unrestricted and adaptive access to all backdoor oracles*. At the core of our results lie new reductions from cryptographic problems to the *communication complexities* of various two-party tasks. Along the way we establish a communication complexity lower bound for set-intersection for cryptographically relevant ranges of parameters and distributions and where set-disjointness can be easy.

Keywords. Random oracle, combiner, communication complexity, set-disjointness, set-intersection, lower bounds.

1 Introduction

Hash functions are one of the most fundamental building blocks in the design of cryptographic protocols. From a provable security perspective, a particularly successful methodology to use hash functions in protocols has been the introduction of the random-oracle (RO) model [14,4]. This model formalizes the intuition that the outputs of a well-designed hash function look random by giving all parties, honest or otherwise, oracle access to a uniformly chosen random function. The strong randomness properties inherent in the oracle, in turn, facilitate the security analyses of many protocols.

The cryptanalytic validation of hash functions can strengthen our confidence in this RO-like behavior. On the other hand, as such analyses improve, (unintentional) weaknesses in hash functions are discovered, which can lead to their

partial or total break of security. However, cryptanalytic validation might also fail to detect *intentional* weaknesses that are built into systems. For example such backdoors might be themselves built using cryptographic techniques, which make them hard to detect. Prominent examples show that such backdoors exist and can be exploited in various ways [5,9,10].

In this work we revisit a classical question on protecting against failures of hash functions. Numerous works in this area have studied if, and to what level, by *combining* different hash functions one can offer such protections; see [6,15,16,19] for theoretical treatments and [29,25,12] for cryptanalytic work. However, most work has their focus on unintentional failures (to protect against cryptanalytic advances). In this work, we consider a more adversarial view of hash function failures and ask if well-designed, but possibly *backdoored* hash functions can be used to build backdoor-free hash functions?

Depending on what well-designed means, what adversarial powers the backdoors provide, and what security goals are targeted, different solutions emerge. Hash-function combiners in the works above typically convert two or more hash functions into a new one that is secure as long as *any* of the underlying hash functions is secure. For example, the concatenation combiner builds a collision-resistant hash function given k hash functions as long as one function is collision resistant. Multi-property combiners for other notions, such as PRG, MAC or PRF security, also exist [16].

Typical combiners, however, do not necessarily offer protection when *all* hash functions fail. Intuitively, the goal here is more challenging as all "sources of hardness" have been rendered useless. Despite this, a number of works [26,19,25,32,22] take a more practical approach and introduce an intermediate *weakened* RO model, where hash functions are vulnerable to strong forms of attack, but are otherwise random.

This is an approach that we also adopt here. Since our goal is to protect against *adversarial* weaknesses (aka. backdoors), we place no assumptions on hash-function weaknesses—they can go well beyond computing random preimages or collisions.

1.1 Contributions

We introduce a substantially weakened RO model where an adversary, on top of hash values, can also obtain *arbitrary functions* of the table of the hash function. We formalize this capability via access to a *backdoor oracle* $\textsc{Bd}(f)$ that on input a function f returns $f(\langle \mathsf{H} \rangle)$, arbitrary auxiliary information about the function table of the hash function H. We call this the *backdoored random-oracle* (BRO) model.

Such backdoors are powerful enough to allow for point inversions—simply hardwire the point y that needs to be inverted into a function $f[y]$ that searches for a preimage of y under H—or finding collisions. But they can go well beyond them. For example, although Liskov [26] proves one-way security of the combiner $\mathsf{H}(0|x_1|x_2)|\mathsf{H}(1|x_2|x_1)$ under *random* inversions, it becomes insecure when inverted points are not assumed to be random: given $y_1|y_2$ simply look for an

inverse $0|x'_1|x'_2$ for y_1 such that $1|x'_2|x'_1$ also maps to y_2. BRO can also model arbitrary *preprocessing* attacks (aka. non-uniform attacks) as any auxiliary information about $\langle \mathsf{H} \rangle$ can be computed via a one-time oracle access at the onset. This means that collisions (without salting) can be easily found. Furthermore, since Bd calls can be adaptive, salting does not help in our setting at all. Indeed, with a single hash function and arbitrary backdoor capabilities no combiner can exist as any construction $\mathsf{C}^{\mathsf{H}}(x)$ can be easily inverted by a function that sees the entire $\langle \mathsf{H} \rangle$ and searches for inversions.

In practice it is natural to assume that independent hash functions are available. We can easily model this by an extension to the k-BRO model, whereby k independent ROs and their respective backdoor oracles are made available.[1] The interpretation in our setting is that different "trusted" authorities have designed and made public hash functions that display good (i.e., RO-like) behaviors, but their respective backdoors enable computing any function of the hash tables. We ask if these hash functions can be combined in way that renders their backdoors useless. We observe that the result of Hoch and Shamir [19] can be seen as one building a collision-resistant hash function in the 2-BRO model assuming backdoor oracles that allow for random inversions only.

From a high-level point of view, our main result shows that in the 2-BRO model cryptographic hardness *can* be bootstrapped, even with access to *both* backdoor oracles and even when *arbitrary* backdoor capabilities are provided. In other words, there are secure constructions in the 2-BRO model that can tolerate arbitrary weaknesses in all underlying hash functions. At the core of our results lies new links with hard problems in the area of *communication complexity*.

Communication complexity. The communication complexity [37,23] of a two-party task $f(S,T)$ is the minimum communication cost over two-party protocols that compute $f(S,T)$. Two rich and well-studied problems in this area are the *set-disjointness* and *set-intersection* problems (see [8] for a survey). Here two parties hold sets S and T respectively. In set-disjointness, their goal is to decide whether or not $S \cap T = \emptyset$; in set-intersection they need to compute at least one element in this intersection. Typically, work in communication complexity studies communication cost over all inputs, that is, the *worst-case* communication complexity of a problem, as the focus is on lower bounds. Cryptographic applications, on the other hand, usually require *average-case* hardness. Distributional (average-case) communication complexity of a problem averages the communication cost over random choices of (S,T) from some distribution μ. We will rely on average-case lower bounds in this work.

The basic ideas. In this work, we focus on three of the most important hash combiners: concatenation, cascade, and xor of hash functions H_1 and H_2:

$$\mathsf{C}_{|}^{\mathsf{H}_1,\mathsf{H}_2}(x) := \mathsf{H}_1(x)|\mathsf{H}_2(x)\ , \qquad \mathsf{C}_{\circ}^{\mathsf{H}_1,\mathsf{H}_2}(x) := \mathsf{H}_2(\mathsf{H}_1(x))\ ,$$

[1] Note that k-BRO can be viewed as a *restricted* version of the 1-BRO model where k ROs are built from a single RO acting on k separate domains and backdoor capabilities are restricted to these domains only.

$$\mathsf{C}_{\oplus}^{\mathsf{H}_1,\mathsf{H}_2}(x) := \mathsf{H}_1(x) \oplus \mathsf{H}_2(x) \ .$$

Here $\mathsf{H}_1 : \{0,1\}^n \to \{0,1\}^{n+s_1}$ and $\mathsf{H}_2 : \{0,1\}^n \to \{0,1\}^{n+s_2}$ in the first construction, $\mathsf{H}_1 : \{0,1\}^n \to \{0,1\}^{n+s_1}$ and $\mathsf{H}_2 : \{0,1\}^{n+s_1} \to \{0,1\}^{n+s_1+s_2}$ in the second construction, and $\mathsf{H}_1, \mathsf{H}_2 : \{0,1\}^n \to \{0,1\}^{n+s}$ in the third.

Consider the one-way security of the concatenation combiner in the 2-BRO model. An adversary is given a point $y^* := y_1^* | y_2^* := \mathsf{H}_1(x^*) | \mathsf{H}_2(x^*)$ for a random x^*. It has access to the backdoor oracles $\textsc{Bd}_1$ and $\textsc{Bd}_2$ for functions H_1 and H_2 respectively. Its goal is to compute a preimage x for y^* under $\mathsf{C}_{|}^{\mathsf{H}_1,\mathsf{H}_2}$. This is the case iff $\mathsf{H}_1(x) = y_1^*$ and $\mathsf{H}_2(x) = y_2^*$. Now define two sets $S := \mathsf{H}_1^-(y_1^*)$, the set of preimages of y_1^* under H_1, and $T := \mathsf{H}_2^-(y_2^*)$, the set of preimages of y_2^* under H_2. *Thus the adversary wins iff* $x \in S \cap T$.

The two backdoor oracles respectively know S and T as they are part of the descriptions of the two hash functions. This allows us to convert a successful one-way adversary to a two-party protocol that computes an element x of the intersection $S \cap T$. Put differently, if the communication complexity of set-intersection for sets that are distributed as above has a high lower bound, then the adversary has to place a large number of queries, which, in turn, allows us to conclude that the concatenation combiner is one-way in the 2-BRO model.

The question is: for which sets S and T is set-intersection hard? Suppose the hash functions $\mathsf{H}_1, \mathsf{H}_2 : \{0,1\}^n \to \{0,1\}^m$ are compressing and $m = n - s$. Then on average the sets S and T would each have 2^s elements. We can of course communicate these sets in $\mathcal{O}(2^s)$ bits and find a preimage. However, the cost of this attack when s is linear in n (or even super-logarithmic in n) becomes prohibitive. This raises the question if set-intersection is hard for, say, $s = n/2$ and where the distribution over (S,T) is induced by the two hash functions, where except a single element in common (guaranteed to exist by the rules of the one-way game) all others are sampled uniformly and independently at random and included in the sets.

We observe that hardness of the set-disjointness problem implies hardness of set-intersection as the parties can verify that a given element is indeed in both their sets.[2] Set-disjointness is a better studied problem. To the best of our knowledge two results on set-disjointness with parameters and distributions close to those in our setting have been proven. First, a classical (and technical) result of Babai, Simon and Frankl [1] which shows an $\Omega(\sqrt{N})$ lower bound for random and independent sets S and T of size *exactly* $\sqrt{N}$ in a universe of size N. Second, a result based on information-theoretic arguments due to Bar-Yossef et al. [2], for dependent sets S and T, which has been adapted to *Bernoulli product distributions* in lectures by Moshkovitz and Barak [31, Lecture 9] and Guruswami and Cheraghchi [18, Lecture 21]. The distribution is as follows: for each of the N elements in the universe, independent $\mathrm{Ber}(1/\sqrt{N})$ bits are sampled. (The probability of 1 is $1/\sqrt{N}$.) The sets then consist of all elements for which

[2] On the other hand, for sufficiently large sets that intersect with high probability, set-disjointness is easy whereas set-intersection can remain hard.

the bit is set to 1.[3] The authors again prove an $\Omega(\sqrt{N})$ lower bound (which is tight up to logarithmic factors). We note that both these results only hold for protocols that err with probability at most $\varepsilon \leq 1/100$. However, we only found incomplete proofs of set-disjointness for product Bernoulli distributions, and thus have included a self-contained proof in Appendix C. We also prove a distributional communication complexity lower bound for set-*intersection* for parameters where set-disjointness can be *easy*.

The second result is better suited for our purposes as the size restriction in the first one would restrict us to regular random oracles. Indeed, the distribution induced on the preimages of y_1^* (resp. y_2^*) by the hash function outside the common random point *is* Bernoulli: $\Pr[\mathsf{H}_1(x) = y_1^*] = 1/2^m$ (resp. $\Pr[\mathsf{H}_2(x) = y_2^*] = 1/2^m$) for any x and independently for values of x. We use this fact to show that set-intersection and set-disjointness problems are, respectively, sufficient to prove it is hard to invert random co-domain points (a property that we call random preimage resistance, rPre) or even decide if a preimage exists (which we call oblivious PRG, oPRG). The main benefit of these games is that they do away with the common point guaranteed to exist by the rules of one-way game (and also similar technicalities associated with the standard PRG game). These games can then be related to the one-way and PRG games via cryptographic reductions.

Our lower bound for set-intersection allows us to prove strong one-way security for some parameters, while the set-disjointness bound only enables proving weak PRG security. Using amplification techniques we can then convert the weak results to strong one-way functions [17] or strong PRGs [28]. Note that the reductions for all these results are fully black-box and thus would relativize [33]. This implies that the same proofs also hold in the presence of backdoor oracles. Construction of other primitives in minicrypt also relativize. This means we also obtain backdoor-free PRFs, MACs, PRPs, and symmetric encryption schemes in our model. The resulting constructions, however, are often too inefficient to be of any practical use. The bottleneck for PRG efficiency here is the proven lower bounds for set-disjointness. New lower bounds that give trade-offs between protocol error and communication complexity will enable more efficient/secure constructions. We discuss in Section 4 why the current proof does not permit this.

Recall that collision resistance can *not* be based on one-way functions [35]. The concatenation combiner, on the other hand, appears to be collision resistant as *simultaneous collisions* seem hard to find, even with respect to arbitrary backdoors for each hash function. Indeed, an analysis of collision resistance for this combiner reveals a natural *multi-instance* analogue of the set-intersection problem, which to the best of our knowledge has not been studied yet. Assuming the hardness of this problem (which we leave open) we get collision resistance. We note that fully black-box amplification for collision-resistance also exists [7],

[3] The expected size of such Bernoulli sets is $N/\sqrt{N} = \sqrt{N}$, but this size can deviate from the mean and this distribution is *not* identical to that by Babai et al. [1].

Combiner	Strong OW	Weak PRG	Strong CR
Concatenation	$s_1, s_2 = -(\epsilon+1)\cdot n/2$ for $0 < \epsilon < 1/3$	$s_1 = -n/2+1$, $s_2 = -n/2$	$s_1, s_2 \leq -n/2-1$ (Conjectural)
Cascade	$s_1 = (1+\epsilon)\cdot n, s_2 = -n$ for $-1/2 < \epsilon < 0$	$s_1 = 2n$, $s_2 = -2n+1$	$s_1 = 2n$, $s_2 = -2n-1$ (Conjectural)
Xor	$s = \epsilon \cdot n$ for $-0.42 < \epsilon < 0$	Any s (Conjectural)	Any s (Conjectural)

Table 1: Overview of results for concatenation, cascade, and xor. Functions H_i have stretch s_i. For xor, $s_1 = s_2 = s$. The parameters for collision resistance and the weak PRG security of xor are conjectural.

and it is sufficient to prove hardness for small values of protocol error ε (should this be the case as in the case single-instance set-disjointness).

We carry out similar analyses for the *cascade* and *xor* combiners, for which different choices of parameters lead to security. Although the overall approach remains the same, we need to deal with some difficulties in the security proofs of these combiners. For the cascade combiner these arise from the fact that one of the sets is the *image* of a hash function. The latter distribution is somewhat different to Bernoulli sets (as elements are not chosen independently). We show, however, that by addition of noise one-way and PRG security can be based on *known* lower bounds. For collision-resistance we give a reduction to a multi-instance analogue of set-intersection (whose hardness remains open). For the xor combiner, the communication complexity problem that directly underlies our reduction for one-wayness is multi-instance: both parties hold multiple sets and need to find an intersection in two of their sets. We are, however, able to relate this problem to the standard set-intersection by generating one larger set for each party which labels and contains all smaller sets.

We summarize our results in Table 1. Roughly speaking, strong security demands that the advantage of adversaries in the corresponding security game is negligible, while for weak security it suffices that the advantage is not overwhelming. In the table, concatenation is with respect to hash function $\mathsf{H}_1 : \{0,1\}^n \to \{0,1\}^{n+s_1}$ and $\mathsf{H}_2 : \{0,1\}^n \to \{0,1\}^{n+s_2}$. Cascade is with respect to hash function $\mathsf{H}_1 : \{0,1\}^n \to \{0,1\}^{n+s_1}$ and $\mathsf{H}_2 : \{0,1\}^{n+s_1} \to \{0,1\}^{n+s_1+s_2}$. The xor combiner is with respect to hash functions $\mathsf{H}_1, \mathsf{H}_2 : \{0,1\}^n \to \{0,1\}^{n+s}$. The stretch values s_1, s_2, and s can assume negative values (compressing), positive values (expanding), or be zero (length-preserving).

1.2 Discussion

Backdoors as weaknesses. One of the main motivations for the works of Liskov [26] and Hoch and Shamir [19] is the study of design principles for symmetric schemes that can offer protections against weaknesses in their underlying primitives. For example, Hoch and Shamir study the failure-friendly double-pipe hash construction of Lucks [27]. Similarly, Liskov shows that his *zipper hash* is indifferentiable from a random oracle even with an inversion oracle for its underlying compression function. Proofs of security in the unrestricted BRO model

would strengthen these results as they place weaker assumptions on the types of weaknesses that are discovered.

Auxiliary inputs. As mentioned above, a closely related model to BRO is the Auxiliary-Input RO (AI-RO) model, introduced by Unruh [36] and recently refined by Dodis, Guo, and Katz [13] and Coretti et al. [11]. Here the result of a one-time preprocessing attack with access to the full table of the random oracle is made available to an adversary. The BRO and AI-RO models are similar in that they both allow for arbitrary functions of the random oracle to be computed. However, BRO allows for adaptive, instance-dependent auxiliary information, whereas the AI-RO model only permits a one-time access at the onset.[4] Thus AI-RO is identical to BRO when only a single Bd query at the onset is allowed. Extension to multiple ROs can also be considered for AI-ROs, where independent preprocessing attacks are performed on the hash functions. A corollary is that any positive result in the k-BRO model would also hold in the k-AI-RO model. Results in k-AI-RO model can be proven more directly using the decomposition of high-entropy densities as the setting is non-interactive.

Feasibility in 1-BRO. As already observed, any combiner in 1-BRO is insecure with respect to arbitrary backdoors. We can, however, consider a model where backdoor capabilities are restricted to inversions only. Security in such models will depend on the exact specification of backdoor functionalities $\mathcal{F}$. For example, under random inversions positive results can be established using standard lazy sampling techniques. But another natural choice is to consider functions which output possibly adversarial preimages, i.e., functions $f[y]$ whose outputs are restricted to those x for which $\mathsf{H}(x) = y$. As we have seen, under such generalized inversions provably secure constructions can fail. Moreover, proving security under general inversions seems to require techniques from communication complexity as we do here.

Other settings. Proofs in the random-oracle model often proceed via direct information-theoretic analyses. Here we give cryptographic reductions (somewhat similarly to the standard model) that isolate the underlying communication complexity problems. These problems have diverse applications in other fields (such as circuit complexity, VLSI design, and combinatorial auctions), which motivate their study outside cryptographic contexts. Any improvement in lower bounds for them would also lead to improvements in the security/efficiency of cryptographic constructions. We discussed the benefits of proofs for arbitrary error above. As other examples, results in multi-party communication complexity would translate to the k-BRO model for $k > 2$ or those in quantum communication complexity can be used to built quantum-secure BRO combiners.

Future work. Our work leaves a number of problems open, some of which are closer to work in communication complexity. Lower bounds for set-disjointness that do not assume a small error would improve the security and/or efficiency of

[4] Arguably, the AI-RO model is better named the Non-Uniform RO model: auxiliary input is often instance dependent whereas non-uniformity is not.

our PRG constructions. Moreover, we do not currently have a lower bound for the multi-instance analogue of set-intersection that we need for proving collision resistance. Finding the "maximal" backdoor capabilities in the 1-BRO model under which hardness can be bootstrapped remains an interesting open problem. Katz, Lucks, and Thiruvengadam [21] study the construction of collision-resistant hash functions from ideal ciphers that are vulnerable to differential related-key attacks. We leave the study of combiners for other backdoored primitives, such as ideal ciphers, for future work.

2 Preliminaries

We let $\mathbb{N}$ denote the set of non-negative integers and $\{0,1\}^n$ be the set of all binary strings of length $n \in \mathbb{N}$. For two bit strings x and y, we denote their concatenation by $x|y$. We let $[N]$ denote the set $\{1,\ldots,N\}$. For a finite set S, we denote by $s \twoheadleftarrow S$ the uniform random variable over S. The Bernoulli random variable $x \twoheadleftarrow \mathrm{Ber}(p)$ takes value 1 with probability p and 0 with probability $1-p$. The Binomial random variable $x_1,\ldots,x_n \twoheadleftarrow \mathrm{Bin}(n,p)$ constitutes a sequence of n independent Bernoulli samples. We will sometimes use $e^{-x} := \lim_{n\to\infty}(1-x/n)^n$.

2.1 Random oracles

A hash function H with n-bit inputs and m-bit outputs is simply a function with signature $\mathsf{H} : \{0,1\}^n \to \{0,1\}^m$. We let $\mathrm{Fun}[n,m]$ denote the set of all such functions. $\mathrm{Fun}[n,m]$ is finite and we endow it with the uniform distribution. For a hash function H, we let $\langle\mathsf{H}\rangle$ denote the function table of H encoded as a string of length $m2^n$. We see the x-th m-bit block of $\langle\mathsf{H}\rangle$ as $\mathsf{H}(x)$, identifying strings $x \in \{0,1\}^n$ with integers in $[1,2^n]$. The random-oracle (RO) model (for a given n and m) is a model of computation where all parties have oracle access to a function $\mathsf{H} \twoheadleftarrow \mathrm{Fun}[n,m]$.

Backdoor functions. A backdoor function for $\mathsf{H} \in \mathrm{Fun}[n,m]$ is a function $f : \mathrm{Fun}[n,m] \to \{0,1\}^t$. A backdoor capability class $\mathcal{F}$ is a set of such backdoor functions. The unrestricted class contains all functions. But the class can be also restricted, for example, functions $f[y]$ for $y \in \{0,1\}^m$ whose outputs x are restricted to be in $\mathsf{H}^-(y)$, where $\mathsf{H}^-(y)$ is the set preimages of y under H. Randomness can also be hardwired.

The BRO model. In the backdoored random-oracle (BRO) model, a random function $\mathsf{H} \twoheadleftarrow \mathrm{Fun}[n,m]$ is sampled. All parties are provided with oracle access to H. Adversarial parties are additionally given access to the procedure

$$\underline{\text{Proc. } \textsc{Bd}(f)}:\ \textbf{return}\ f(\langle\mathsf{H}\rangle)$$

for $f \in \mathcal{F}$. Formally, we denote this model by $\mathrm{BRO}[n,m,\mathcal{F}]$, but will omit $[n,m,\mathcal{F}]$ when it is clear from the context. When $\mathcal{F} = \emptyset$, we recover the conventional RO model. As discussed in the introduction, when the adversarial parties

Game $\mathrm{OW}^{\mathcal{A}}_{\mathsf{C}}$	Game $\mathrm{PRG}^{\mathcal{A}}_{\mathsf{C}}$	Game $\mathrm{CR}^{\mathcal{A}}_{\mathsf{C}}$
for $i = 1, 2$ **do**	**for** $i = 1, 2$ **do**	**for** $i = 1, 2$ **do**
$\quad \mathsf{H}_i \twoheadleftarrow \mathrm{Fun}[n_i, m_i]$	$\quad \mathsf{H}_i \twoheadleftarrow \mathrm{Fun}[n_i, m_i]$	$\quad \mathsf{H}_i \twoheadleftarrow \mathrm{Fun}[n_i, m_i]$
$x \twoheadleftarrow \{0,1\}^n;\ y \leftarrow \mathsf{C}^{\mathsf{H}_i}(x)$	$y_0 \twoheadleftarrow \{0,1\}^m;\ b \twoheadleftarrow \{0,1\}$	$(x_1, x_2) \leftarrow \mathcal{A}^{\mathsf{H}_i, \mathrm{BD}_i}$
$x' \twoheadleftarrow \mathcal{A}^{\mathsf{H}_i, \mathrm{BD}_i}(y)$	$x \twoheadleftarrow \{0,1\}^n;\ y_1 \leftarrow \mathsf{C}^{\mathsf{H}_i}(x)$	$y_1 \leftarrow \mathsf{C}^{\mathsf{H}_i}(x_1);\ y_2 \leftarrow \mathsf{C}^{\mathsf{H}_i}(x_2)$
return $(\mathsf{C}^{\mathsf{H}_i}(x') = y)$	$b' \leftarrow \mathcal{A}^{\mathsf{H}_i, \mathrm{BD}_i}(y_b)$	**return** $(x_1 \neq x_2 \wedge y_1 = y_2)$
	return $(b' = b)$	

Fig. 1: The one-way, pseudorandomness, and collision resistance games for a function $\mathsf{C}^{\mathsf{H}_i} \in \mathrm{Fun}[n, m]$.

call the backdoor oracle only once and before any hash queries, we recover random oracles with auxiliary input, the AI-RO model [11, Definition 2]. Thus, BRO also models oracle-dependent auxiliary input or pre-computation attacks as special cases. In the k-BRO model (with the implicit parameters $[n_i, m_i, \mathcal{F}_i]$ for $i = 1, \ldots, k$) access to k independent random oracles $\mathsf{H}_i \in \mathrm{Fun}[n_i, m_i]$ and their respective backdoors BD_i with capabilities $\mathcal{F}_i$ are provided. That is, procedure $\mathrm{BD}_i(f)$ returns $f(\langle \mathsf{H}_i \rangle)$. In this work we are primarily interested in the 1-BRO and 2-BRO models with *unrestricted* $\mathcal{F}$.

We observe that the 2-BRO$[n, m, \mathcal{F}_1, n, m, \mathcal{F}_2]$ model is identical to the 1-BRO$[n+1, m, \mathcal{F}]$ model where for $\mathsf{H} \in \mathrm{Fun}[n+1, m]$ we define $\mathsf{H}_1(x) := \mathsf{H}(0|x)$, $\mathsf{H}_2(x) := \mathsf{H}(1|x)$ and $\mathcal{F}$ to consist of two types of functions: those in $\mathcal{F}_1$ and dependent on values $\mathsf{H}(0|x)$, that is the function table of H_1, only, and those in $\mathcal{F}_2$ and dependent on values of $\mathsf{H}(1|x)$, that is the function table of H_2, only. Thus the adversary in the unrestricted 2-BRO model has less power than in the unrestricted 1-BRO model.

2.2 Cryptographic notions

We recall the basic notions of one-wayness, pseudorandomness, and collision-resistance for a construction $\mathsf{C}^{\mathsf{H}_1, \mathsf{H}_2}$ in the 2-BRO model in Figure 1. We omit the implicit parameters from the subscripts and use $\mathsf{C}^{\mathsf{H}_i}$ in place of $\mathsf{C}^{\mathsf{H}_1, \mathsf{H}_2}$ to ease notation. These notions can also be defined in the 1-BRO model analogously by removing access to H_2 and BD_2 throughout. The advantage terms are

$$\mathsf{Adv}^{\mathrm{ow}}_{\mathsf{C}^{\mathsf{H}_i}}(\mathcal{A}) := \Pr[\mathrm{OW}^{\mathcal{A}}_{\mathsf{C}^{\mathsf{H}_i}}], \qquad \mathsf{Adv}^{\mathrm{prg}}_{\mathsf{C}^{\mathsf{H}_i}}(\mathcal{A}) := 2 \cdot \Pr[\mathrm{PRG}^{\mathcal{A}}_{\mathsf{C}^{\mathsf{H}_i}}] - 1,$$

$$\mathsf{Adv}^{\mathrm{cr}}_{\mathsf{C}^{\mathsf{H}_i}}(\mathcal{A}) := \Pr[\mathrm{CR}^{\mathcal{A}}_{\mathsf{C}^{\mathsf{H}_i}}] \ .$$

All probabilities in this model are also taken over random choices of H_i. Informally $\mathsf{C}^{\mathsf{H}_1, \mathsf{H}_2}$ is OW, PRG, or CR if the advantage of any adversary $\mathcal{A}$ querying its oracles, such that the total length of the received responses remains "reasonable", is "small". Note that if one only considers backdoor functions with 1-bit

Game $\mathrm{rPre}_{\mathsf{C}}^{\mathcal{A}}$	Game $\mathrm{oPRG}_{\mathsf{C}}^{\mathcal{A}}$	Game $\mathrm{IU}_{\mathsf{C}}^{\mathcal{A}}$
for $i = 1, 2$ **do**	**for** $i = 1, 2$ **do**	**for** $i = 1, 2$ **do**
$\quad \mathsf{H}_i \twoheadleftarrow \mathrm{Fun}[n_i, m_i]$	$\quad \mathsf{H}_i \twoheadleftarrow \mathrm{Fun}[n_i, m_i]$	$\quad \mathsf{H}_i \twoheadleftarrow \mathrm{Fun}[n_i, m_i]$
$y \twoheadleftarrow \{0,1\}^m$	$y \twoheadleftarrow \{0,1\}^m$	$y_1 \twoheadleftarrow \mathsf{Img}(\mathsf{C}^{\mathsf{H}_i});\ b \twoheadleftarrow \{0,1\}$
$x' \leftarrow \mathcal{A}^{\mathsf{H}_i, \mathrm{BD}_i}(y)$	$b' \leftarrow \mathcal{A}^{\mathsf{H}_i, \mathrm{BD}_i}(y)$	$x \twoheadleftarrow \{0,1\}^n;\ y_0 \leftarrow \mathsf{C}^{\mathsf{H}_i}(x)$
if $y \in \overline{\mathsf{Img}}(\mathsf{C}^{\mathsf{H}_i})$	**return** $(b' = (y \in \mathsf{Img}(\mathsf{C}^{\mathsf{H}_i})))$	$b' \leftarrow \mathcal{A}^{\mathsf{H}_i, \mathrm{BD}_i}(y_b)$
$\quad$ **return** $(x' = \bot)$		**return** $(b' = b)$
return $(\mathsf{C}^{\mathsf{H}_i}(x') = y)$		

Fig. 2: The random preimage resistance (rPre), oblivious PRG (oPRG), and image uniformity (IU) games for $\mathsf{C}^{\mathsf{H}_i} \in \mathrm{Fun}[n, m]$.

output lengths, the total length of the oracle responses directly translates to the number of queries made by $\mathcal{A}$. We denote by $\mathsf{Q}(\mathcal{A})$ the number of oracle queries made by an adversary $\mathcal{A}$ to H_i and BD_i. Weak security in each case means that the corresponding advantage is less than 1 and not overwhelming.

We define variants of the above games which will be helpful in our analyses. For a function $\mathsf{H} \in \mathrm{Fun}[n, m]$, define $\mathsf{Img}(\mathsf{H}) := \mathsf{H}(\{0,1\}^n)$ and $\overline{\mathsf{Img}}(\mathsf{H}) := \{0,1\}^m \setminus \mathsf{Img}(\mathsf{H})$. The *random preimage-resistance* (rPre) game is defined similarly to everywhere preimage-resistance (ePre) [34] except that a random co-domain point (as opposed to any such point) must be inverted. This definition differs from one-way security in two aspects: the distribution of $\mathsf{H}(x)$ for a uniform x might not be uniform. Furthermore, some points in the co-domain might not have any preimages. We also define a decisional variant, called *oblivious PRG* (oPRG), where the adversary has to decide if a random co-domain point has a preimage. We formalize these games in Figure 2. The advantage terms are defined as:

$$\mathsf{Adv}^{\mathrm{rpre}}_{\mathsf{C}^{\mathsf{H}_i}}(\mathcal{A}) := \Pr[\mathrm{rPre}^{\mathcal{A}}_{\mathsf{C}^{\mathsf{H}_i}}] \qquad \mathsf{Adv}^{\mathrm{oprg}}_{\mathsf{C}^{\mathsf{H}_i}}(\mathcal{A}) := \Pr[\mathrm{oPRG}^{\mathcal{A}}_{\mathsf{C}^{\mathsf{H}_i}}]$$

Weak analogues of the above security notions (for example weak rPre or weak oPRG) are defined by requiring the advantage to be bounded away from 1 (i.e., not to be overwhelming). These definitions can be formalized in the asymptotic language, but we use concrete parameters here.

We state two lemmas that relate OW and rPre, resp. PRG and oPRG: for functions that have *uniform images*, as defined below, we show that OW security is implied by rPre security and PRG security is implied by oPRG security.

Image Uniformity. Let $\mathsf{C}^{\mathsf{H}_i} \in \mathrm{Fun}[n, m]$ be a construction in the 2-BRO model. In the image uniformity game IU defined in Figure 2, an adversary, given access to all backdoor oracles, must decide whether a given value is chosen uniformly at random from the image of $\mathsf{C}^{\mathsf{H}_i}$ or computed as the image of a value

x chosen uniformly at random from the domain. The advantage term is

$$\mathsf{Adv}^{\mathrm{iu}}_{\mathsf{C}^{\mathsf{H}_i}}(\mathcal{A}) := 2 \cdot \Pr[\mathrm{IU}^{\mathcal{A}}_{\mathsf{C}^{\mathsf{H}_i}}] - 1 \ ,$$

where the probability is taken over random choices of H_i.

The following lemma upper bounds the advantage of adversaries playing the image uniformity game for combiners with different stretch values. We denote by $\mathcal{U}_S$ the uniform distribution over a set S. We also let $\mathcal{U}^{\mathsf{p}}_f$ denote the distribution defined by $\mathcal{U}^{\mathsf{p}}_f(x) = |f^{-1}(x)|/2^n$, where $f \in \mathrm{Fun}[n,m]$ is a uniform function. We refer the readers to Appendix A for the proofs.

Lemma 1 (Combiner image uniformity). *Let $\mathsf{C}^{\mathsf{H}_1,\mathsf{H}_2}_t : \{0,1\}^n \to \{0,1\}^m$ be a combiner for $t \in \{|, \circ, \oplus\}$. Let $\mathsf{H} : \{0,1\}^n \to \{0,1\}^m$ be a hash function. Then*

$$\mathsf{Adv}^{\mathrm{iu}}_{\mathsf{C}^{\mathsf{H}_i}_t}(\mathcal{A}) \leq \mathbb{E}_{\mathsf{H}}\left[\Delta_{\mathsf{TV}}\left(\mathcal{U}_{\mathsf{Img}(\mathsf{H})}, \mathcal{U}^{\mathsf{p}}_{\mathsf{H}}\right)\right] + 2 \cdot p_t \ ,$$

where $p_| = p_\oplus = 0$ and $p_\circ \leq 2^{2n_1 - m_1}$ is the probability that $\mathsf{H}_1 : \{0,1\}^{n_1} \to \{0,1\}^{m_1}$ is not injective (i.e., it has at least one collision). Let $2^n = C \cdot 2^{m \cdot \gamma}$ for constants C and γ. Then the above statistical distance is negligible for $\gamma > 1$ and $0 < \gamma < 1$ when $C = 1$, while for $\gamma = 1$ and $C \leq 1$ it less than $e^{-C} \cdot \left(C/(1 - e^{-C}) - 1\right)$ plus negligible terms.

Now we can relate our notions of rPre and oPRG with their classical variants, i.e., one-way and PRG security. Proofs of both lemmas are included in Appendix B.

Lemma 2 (rPre + IU $\implies$ OW). *Let $\mathsf{C}^{\mathsf{H}_i} \in \mathrm{Fun}[n,m]$ be a construction in the 2-BRO model. Then for any adversary $\mathcal{A}$ against the one-way security of $\mathsf{C}^{\mathsf{H}_i}$, there is an adversary $\mathcal{B}$ against the image uniformity and an adversary $\mathcal{C}$ against the* rPre *security of $\mathsf{C}^{\mathsf{H}_i}$, all in the 2-BRO model and using identical backdoor functionalities, such that*

$$\mathsf{Adv}^{\mathrm{ow}}_{\mathsf{C}^{\mathsf{H}_i}}(\mathcal{A}) \leq \mathsf{Adv}^{\mathrm{iu}}_{\mathsf{C}^{\mathsf{H}_i}}(\mathcal{B}) + \frac{1}{\alpha} \cdot \mathsf{Adv}^{\mathrm{rpre}}_{\mathsf{C}^{\mathsf{H}_i}}(\mathcal{C}) - \frac{1-\alpha}{\alpha} \ ,$$

where $\alpha := \Pr[y \in \mathsf{Img}(\mathsf{C}^{\mathsf{H}_i})]$ over a random choice of $y \in \{0,1\}^m$ and H_i.

An analogous result also holds for oPRG security.

Lemma 3 (oPRG + IU $\implies$ PRG). *Let $\mathsf{C}^{\mathsf{H}_i} \in \mathrm{Fun}[n,m]$ be a construction in the 2-BRO model which is expanding with $m - n \geq 0.53$. Then for any adversary $\mathcal{A}$ against the PRG security of $\mathsf{C}^{\mathsf{H}_i}$, there is an adversary $\mathcal{B}$ against the image uniformity and an adversary $\mathcal{C}$ against the* oPRG *security of $\mathsf{C}^{\mathsf{H}_i}$, both in the 2-BRO model and using identical backdoor functionalities, such that*

$$\mathsf{Adv}^{\mathrm{prg}}_{\mathsf{C}^{\mathsf{H}_i}}(\mathcal{A}) \leq \mathsf{Adv}^{\mathrm{iu}}_{\mathsf{C}^{\mathsf{H}_i}}(\mathcal{B}) + \frac{1-\alpha}{\alpha} \cdot \mathsf{Adv}^{\mathrm{oprg}}_{\mathsf{C}^{\mathsf{H}_i}}(\mathcal{C}) - (1-\alpha) \ ,$$

where $\alpha := \Pr[y \in \mathsf{Img}(\mathsf{C}^{\mathsf{H}_i})]$ over a random choice of $y \in \{0,1\}^m$ and H_i.

3 Black-Box Combiners

A standard way to build a good hash function from a number of possibly "faulty" hash functions is to combine them [24]. For instance, given k hash functions $\mathsf{H}_1, \ldots, \mathsf{H}_k$, the classical concatenation combiner is guaranteed to be collision resistant as long as one out of the k hash functions is collision resistant. More formally, a black-box collision-resistance combiner C is a pair of oracle circuits $(\mathsf{C}^{\mathsf{H}_i}, \mathsf{R}^{\mathcal{A}})$ where $\mathsf{C}^{\mathsf{H}_i}$ is the construction and $\mathsf{R}^{\mathcal{A}}$ is a reduction that given as oracle any procedure $\mathcal{A}$ that finds a collision for $\mathsf{C}^{\mathsf{H}_i}$, returns collisions for *all* of the underlying H_i's. We are interested in a setting where *none* of the available hash functions is good. Under this assumption, however, a secure hash function must be built from scratch, implying that the source of cryptographic hardness must lie elsewhere. As we discussed above, this question has been studied in the RO model.

We briefly explore the difficulty in the standard model here. We consider a variant of this problem where the hash functions are weak due to the existence of *backdoors*. A generation algorithm Gen outputs keys (hk, bk), where hk is used for hashing and bk enables an unspecified backdoor capability (such as finding preimages or collisions). Our hardness assumption is that the hash function with key hk is collision resistant without access to bk. However, when bk is available, no security is assumed. In this setting, the definition of a combiner can be simplified: instead of requiring the existence of a reduction $\mathsf{R}^{\mathcal{A}}$ as above, we can proceed in the standard way and require that the advantage of any adversary $\mathcal{A}(S)$ that gets any subset $S \subset \{bk_1, \ldots, bk_k\}$ of the backdoors of size $|S| \leq k-1$ to be small.[5] Let us call a combiner secure against any set of at most t backdoors a $\binom{k}{k-t}$-combiner.

It is trivial to see that a $\binom{k}{0}$-combiner is also a $\binom{k}{1}$-combiner. It is also easy to see that a black-box combiner is also a $\binom{k}{1}$-combiner. We are, however, interested in the feasibility of $\binom{k}{0}$-combiners. In this setting there *is* an assumed source of hardness, namely the collision resistance of hash functions without backdoors. But constructions that have to work with a *provided* set of keys seem hard.[6] We next give a simple impossibly result that formalizes this intuition under fully black-box constructions.

Theorem 1. *For any positive $k \in \mathbb{N}$, there are no fully black-box constructions of compressing collision-resistant $\binom{k}{0}$-combiners.*

[5] The classical setting can be viewed as one where bk's are *fixed*, which leads to a difficulty when the new definition is used: a combiner (formally speaking) can "detect" which hash functions are the good ones and use them. Since this detection procedure is not considered practical, one instead asks for the existence of a reduction R as discussed above.

[6] Without this restriction, a trivial construction exists: generate a fresh hash key and "forget" its backdoor. In practice, however, hash keys model sampling of a (unkeyed) hash function from a family. Moreover, it is unclear if the designer of the combiner will securely erase the generated backdoor. Thus, we assume that for any generated key its backdoor is also available.

Proof idea. Let $(\mathsf{H}, \mathcal{A})$ be a pair of oracles such that $\mathsf{H}(hk, \cdot)$ implements a random function and $\mathcal{A}(\langle \mathsf{C} \rangle, hk_1, \ldots, hk_k, bk_1, \ldots, bk_k)$ is a break oracle that operates as follows. It interprets $\langle \mathsf{C} \rangle$ as the description of a combiner. It then checks that each bk_i indeed enables generating collisions under hk_i. If so, it (inefficiently) finds a random collision for $\mathsf{C}^{\mathsf{H}(hk_1,\cdot),\ldots,\mathsf{H}(hk_k,\cdot)}$ and returns it. An efficient reduction R is given oracle access to $\mathcal{A}$ and H as well as a key hk^* (without its backdoor bk^*). It should find a collision for $\mathsf{H}(hk^*, \cdot)$ while making a small (below birthday) number of queries to the two oracles $\mathcal{A}$ and H. We show that any such reduction R must have a negligible success probability.

We distinguish between two cases based on whether the reduction R uses the provided break oracle $\mathcal{A}$ or not. Without the use of $\mathcal{A}$, the reduction would break collision resistance for hk^* on its own, contradicting the collision resistance of hk^* beyond the birthday bound. To use $\mathcal{A}$ the reduction has to provide it with k keys hk_i and some other keys bk_i that enable finding collisions (since $\mathcal{A}$ checks this). However, none of the provided keys hk_i can be hk^*, since R must also provide some $\tilde{bk}^*$ that enables finding collisions under hk^*, which means that R can directly use $\tilde{bk}^*$ to compute a collision for $\mathsf{H}(hk^*, \cdot)$, once again contradicting the assumed collision resistance of hk^* beyond the birthday bound. Thus, R does not use hk^*. A random oracle $\mathsf{H}(hk^*, \cdot)$, however, is collision resistant even in the presence of random collisions for $\mathsf{H}(hk, \cdot)$ for $hk \neq hk^*$. This means that R, which places a small number of oracle queries, will have a negligible success probability. □

There is room to circumvent this result by considering non-black-box constructions. Here, we will study hash function combiners in the k-BRO model, where the hash oracles model access to different hk and the backdoor oracles model access to the corresponding bk's. As mentioned above, this approach has also been adapted in a number of previous works, both from a provable security as well as a cryptanalytic view [19,30,22,25,20]. In this work we will focus on basic security properties of the concatenation, cascade and xor combiners in the *unrestricted* 2-BRO model.

4 Communication Complexity

The communication cost [37,23] of a two-party deterministic protocol π on inputs (x, y) is the number of bits that are transmitted in a run of the protocol $\pi(x, y)$. We denote this by $\mathsf{CC}(\pi(x, y))$. The worst-case communication complexity of π is $\max_{(x,y)} \mathsf{CC}(\pi(x, y))$. A protocol π computes a task (function) $f : X \times Y \to Z$ if the last message of $\pi(x, y)$ is $f(x, y)$. The communication complexity of a task f is the minimum communication complexity of any protocol π that computes f. Protocols can also be randomized and thus might err with probability $\Pr[\pi(x, y) \neq f(x, y)]$. Following cryptographic conventions, we denote protocol correctness by $\mathsf{Adv}^{\mathrm{f}}_{\mu}(\pi)$, where f is a placeholder for the name of the task f.

In the cryptographic setting we are interested in distributional (aka. average-case) communication complexity measured by averaging the communication cost

over random choices of inputs and coins. A standard coin-fixing argument shows that in the *distributional* setting any protocol can be derandomized with no change in communication complexity, and thus we can focus on deterministic protocols. For a given distribution μ over the inputs (x, y), the protocol error and correctness are computed by taking the probability over the choice of (x, y). We define the distributional communication cost of a deterministic protocol π as

$$\mathsf{D}_\mu(\pi) := \mathbb{E}_{(x,y)\sim\mu}[\mathsf{CC}(\pi(x, y))] \ .$$

The distributional communication complexity of a task f with error ε is

$$\mathsf{D}^\varepsilon_\mu(f) := \min_\pi \mathsf{D}_\mu(\pi) \ ,$$

where the minimum is taken over all deterministic protocols π which err with probability at most ε. In this work, we need to slightly generalize functional tasks to relational tasks $R(x, y) \subseteq Z$ and define error as $\Pr[\pi(x, y) \notin R(x, y)]$.

Two central problems in communication complexity that have received substantial attention are the *set-disjointness* and the *set-intersection* problems. In set-disjointness two parties, holding sets S and T respectively, compute the binary function $\mathrm{DISJ}(S, T) := (S \cap T = \emptyset)$. In set-intersection, their goal is to compute the relation $\mathrm{INT}(S, T) := S \cap T$; that is, the last message of the protocol should be equal to some element in the intersection. Note that set-disjointness can be seen as a decisional version of set-intersection and is easier. As mentioned, we are interested in average-case lower bounds for these tasks and moreover we focus on *product* distributions, where the sets are chosen independently.

Two main results to this end have been proven.[7] A classical result of Babai, Frankl, and Simon [1] establishes an $\Omega(\sqrt{N})$ lower bound for set-disjointness where the input sets S and T are independent random subsets of $[N]$ of size exactly $\sqrt{N}$. This result, however, is restrictive for us as it roughly translates to regular functions in the cryptographic setting. Moreover, its proof uses intricate combinatorial arguments, which are somewhat hard to work with.

A second result considers the following distribution. Each element $x \in [N]$ is thrown into S independently with probability p. (And similarly for T with probability q.) We can view S as a N-bit string X where its i-th bit x_i is 1 iff $i \in S$. Thus the distribution can be viewed as N i.i.d. Bernoulli random variables $x_i \sim \mathrm{Ber}(p)$ where $p := \Pr[x_i = 1]$. Thus the elements of the sets form a binomial distribution, and accordingly we write $S \sim \mathrm{Bin}(N, p)$ and $T \sim \mathrm{Bin}(N, q)$. We define $\mu(p, q)$ as the *product* of these distributions. When $p = q = 1/2$ we get the product uniform distribution over the subsets of $[N] \times [N]$, but typically we will be looking at much smaller values of p and q of order $1/\sqrt{N}$.

Using information-theoretic techniques [2], the following lower bound can be established.

[7] We note that most of the work on distributional communication complexity is driven by Yao's min-max lemma, which lower bounds worst-case communication complexity using distributional communication complexity for *some* (often non-uniform) distribution.

Theorem 2 (Set-disjointness lower bound). *Let $N \in \mathbb{N}$ and assume $p, q \in (0, 1/2]$ with $p \leq q$ and $pq = 1/(\delta N)$ for some $\delta > 1$. Let $\mu(p, q)$ be the product binomial distribution over subsets $S, T \subseteq [N]$. Assume $\varepsilon < \frac{(\delta-1)p_0}{(4+\delta)}$ and let $p_0 := \Pr[\mathrm{DISJ}(S, T) = 0]$. Then*

$$\mathsf{D}^{\varepsilon}_{\mu(p,q)}(\mathrm{DISJ}) \geq \frac{Np}{8} \cdot \big((\delta - 1)p_0 - (4 + \delta)\varepsilon\big)^2 \ .$$

We have included a detailed proof of the above theorem in Appendix C, which follows those in [31,18]. Our proof generalizes the original result, which was only claimed for $p = q = 1/\sqrt{N}$.[8] Roughly speaking, the proof proceeds along the following lines. We can lower bound the communication complexity of any protocol by the total information leaked by its transcripts about each coordinate (x_i, y_i). The latter can be lower bounded based on the statistical distance in protocol transcripts when $x_i = 1 \wedge y_i = 0$ and $x_i = 0 \wedge y_i = 1$. This step uses a number of information-theoretic inequalities, which we include with proofs in Appendix C. Finally, we show that a highly correct protocol can be used as a distinguisher with constant advantage: When $x_i = 0 \wedge y_i = 0$, for a constant fraction of the inputs the sets will be disjoint. However, when $x_i = 1 \wedge y_i = 1$ they necessarily intersect, but this condition happens for a constant fraction of the inputs. We get a $\sqrt{N}$ lower bound by averaging over the i's.

In this section we also prove a communication complexity lower bound for the set-intersection problem over Bernoulli sets for which set-disjointness can be *easy*. Although the overall proof structure will be similar to that in [31,18], we will differ in a number of places. First, as above we leave the Bernoulli parameters free so as to be able to compute a feasible region where the lower bound will be non-trivial. We also use the fact that a candidate element can be checked to belong to the intersection (whereas a decision bit for disjointness cannot be checked for correctness). This ensures that the protocol error is one-sided, and allows us to remove the requirement of ε being sufficiently small. Finally, we will bound the probability that the protocol outputs a *random element* in the intersection. This leads to a distinguisher that succeeds with smaller advantage, but overall will lead to a non-trivial bound. We state and prove the formal result next.

Theorem 3 (Set-intersection lower bound). *Let $N \in \mathbb{N}$ and assume $p, q \in (0, 1/2]$ with $p \leq q$. Let $\mu(p, q)$ be the product binomial distribution over subsets $S, T \subseteq [N]$. Let ε be the protocol error and set $p_0 := \Pr[\mathrm{DISJ}(S, T) = 0]$. If $\varepsilon \leq p_0$ then*

$$\mathsf{D}^{\varepsilon}_{\mu(p,q)}(\mathrm{INT}) \geq \frac{Np}{8} \cdot \Big(\frac{p_0 - \varepsilon}{Npq}\Big)^2 \ .$$

For sufficiently large N we have $p_0 = 1 - (1 - pq)^N \approx 1 - e^{-Npq}$. If $pq \gg 1/N$ we get that $p_0 \approx 1$ (the sets intersect with overwhelming probability) and for the theorem we would need that $\varepsilon \leq 1$.

[8] In Appendix C.3 we give a new refined proof that extends the theorem to $\delta \geq 0.8$.

Let us first give some preliminaries and state two lemmas that are used in the proof of Theorem 3. For random variables X and Y, their statistical distance (aka. total variance) is denoted by $\Delta_{\mathsf{TV}}(X,Y)$, their mutual information is denoted by $I(X;Y)$, and their Hellinger distance is denoted by $\Delta_{\mathsf{Hel}}(X,Y)$:

$$\Delta_{\mathsf{Hel}}(X,Y) := \sqrt{1 - \sum_{z \in D} \sqrt{\Pr[X=z]\Pr[Y=z]}} \ .$$

Statistical and Hellinger distance are related (cf. proofs in Appendix C.1) via:

$$\Delta^2_{\mathsf{Hel}}(X,Y) \leq \Delta_{\mathsf{TV}}(X,Y) \leq \sqrt{2} \cdot \Delta_{\mathsf{Hel}}(X,Y) \ .$$

Below, Lemma 4, proven in Appendix C.1, relates the mutual information of two random variables with their Hellinger distance.

Lemma 4 (Information to Hellinger). *Let X and Y be random variables and $Y_x := Y|X=x$, i.e., Y conditioned on $X=x$. Then*

$$\mathbb{E}_{x \in X}[\Delta^2_{\mathsf{Hel}}(Y, Y_x)] \leq I(X;Y) \ .$$

Next we state the cut-and-paste lemma from communication complexity. A proof is included in Appendix C.2.

Lemma 5 (Cut-and-Paste). *Let $\Pi(X,Y)$ denote a random variable for the transcripts of a deterministic protocol on input bit strings (X,Y) such that the corresponding sets S and T are drawn from μ, i.e., $S,T \sim \mu$. Let $a,b \in \{0,1\}$ and define $\Pi^i_{a,b}(X,Y) := \Pi(X,Y) \mid x_i = a \wedge y_i = b$. Then for each i, it holds that*

$$\Delta^2_{\mathsf{Hel}}(\Pi^i_{0,0}, \Pi^i_{1,1}) = \Delta^2_{\mathsf{Hel}}(\Pi^i_{0,1}, \Pi^i_{1,0}) \ .$$

Now we can prove the claimed lower bound on the communication complexity of set-intersection.

Proof of Theorem 3. Let π be a deterministic protocol with error at most ε, i.e.,

$$\Pr_{(S,T)\sim\mu}[\pi(X,Y) \in \mathrm{INT}(S,T)] \geq 1 - \varepsilon \ ,$$

where X and Y are bit string representations of S and T as explained above. Let $\Pi(X,Y)$ denote a random variable for the transcripts of protocol π on inputs (X,Y) with corresponding sets $(S,T) \sim \mu$. We write $X = (x_1,\ldots,x_N)$ and $Y = (y_1,\ldots,y_N)$ where $x_i, y_i \in \{0,1\}$. For random variables A and B, let $\mathrm{supp}(A)$ denote the support of A (i.e., the set of values that have a non-zero probability of happening), and $H(A)$ denote the Shannon entropy. We have

$$\begin{aligned}
\mathsf{D}^{\varepsilon}_{\mu(p,q)}(\mathrm{INT}) &\geq \log|\mathrm{supp}(\Pi(X,Y))| \\
&\geq H(\Pi(X,Y)) = H(\Pi(X,Y)) + H(X,Y) - H(X,Y,\Pi(X,Y)) \\
&= I(X,Y;\Pi) = I(x_1,\ldots,x_N,y_1,\ldots,y_N;\Pi) \geq \sum_{i=1}^{N} I(x_i,y_i;\Pi) \ ,
\end{aligned}$$

where the last inequality holds due to the independence of $x_1, \ldots, x_N, y_1, \ldots, y_N$ (cf. Lemma 14 and its proof in Section C.1). Let $\Pi^i_{a,b}$ be Π conditioned on the i-th coordinates of X and Y being fixed to a and b respectively:

$$\Pi^i_{a,b}(X,Y) := \Pi(X,Y) \mid x_i = a \wedge y_i = b \ .$$

By Lemma 4 we know

$$I(x_i, y_i; \Pi) \geq \mathbb{E}_{(a,b)}[\Delta^2_{\mathsf{Hel}}(\Pi^i_{a,b}, \Pi)] \ ,$$

where $(a,b) \sim \mathrm{Ber}(p) \times \mathrm{Ber}(q)$ and Δ_{Hel} is the Hellinger distance.

Since $q \geq p$ we have that $q(1-p) \geq p(1-q)$ and since $q \leq 1/2$, we also have that $p(1-q) \geq p/2$. Thus

$$\begin{aligned}
I(x_i, y_i; \Pi) &\geq p(1-q) \cdot \Delta^2_{\mathsf{Hel}}(\Pi^i_{1,0}, \Pi) + q(1-p) \cdot \Delta^2_{\mathsf{Hel}}(\Pi^i_{0,1}, \Pi) \\
&\geq p(1-q) \cdot \left(\Delta^2_{\mathsf{Hel}}(\Pi^i_{1,0}, \Pi) + \Delta^2_{\mathsf{Hel}}(\Pi^i_{0,1}, \Pi)\right) \\
&\geq p/2 \cdot \left(\Delta^2_{\mathsf{Hel}}(\Pi^i_{1,0}, \Pi) + \Delta^2_{\mathsf{Hel}}(\Pi^i_{0,1}, \Pi)\right) \\
&\geq p/4 \cdot \left(\Delta_{\mathsf{Hel}}(\Pi^i_{1,0}, \Pi) + \Delta_{\mathsf{Hel}}(\Pi^i_{0,1}, \Pi)\right)^2 \\
&\geq p/4 \cdot \Delta^2_{\mathsf{Hel}}(\Pi^i_{1,0}, \Pi^i_{0,1}) \ .
\end{aligned}$$

The last inequality is by the triangle inequality for the metric Δ_{Hel}, and the penultimate inequality uses $x^2 + y^2 \geq (x+y)^2/2$. Hence,

$$\begin{aligned}
\mathsf{D}^{\varepsilon}_{\mu(p,q)}(\mathrm{INT}) &\geq N \cdot \mathbb{E}_i[I(x_i, y_i; \Pi)] \\
&\geq Np/4 \cdot \mathbb{E}_i[\Delta^2_{\mathsf{Hel}}(\Pi^i_{1,0}, \Pi^i_{0,1})] \\
&= Np/4 \cdot \mathbb{E}_i[\Delta^2_{\mathsf{Hel}}(\Pi^i_{0,0}, \Pi^i_{1,1})] \\
&\geq Np/8 \cdot \mathbb{E}_i[\Delta^2_{\mathsf{TV}}(\Pi^i_{0,0}, \Pi^i_{1,1})] \\
&\geq Np/8 \cdot \left(\mathbb{E}_i[\Delta_{\mathsf{TV}}(\Pi^i_{0,0}, \Pi^i_{1,1})]\right)^2 \ ,
\end{aligned}$$

where the third inequality uses the cut-and-paste lemma of communication complexity (Lemma 5) which states that $\Delta^2_{\mathsf{Hel}}(\Pi^i_{1,0}, \Pi^i_{0,1}) = \Delta^2_{\mathsf{Hel}}(\Pi^i_{0,0}, \Pi^i_{1,1})$ for any deterministic protocol π. The penultimate inequality uses $\Delta_{\mathsf{TV}}(A,B) \leq \sqrt{2}\Delta_{\mathsf{Hel}}(A,B)$, which implies $\Delta^2_{\mathsf{Hel}}(\Pi^i_{0,0}, \Pi^i_{1,1}) \geq 1/2\Delta^2_{\mathsf{TV}}(\Pi^i_{0,0}, \Pi^i_{1,1})$, and the last inequality is by Jensen. Thus it remains to lower bound $\Delta_{\mathsf{TV}}(\Pi^i_{0,0}, \Pi^i_{1,1})$.

For every i we have

$$\Pr[\Pi^i_{0,0}(X,Y) = i] = 0 \ .$$

This is because we have conditioned on $x_i = y_i = 0$ and the two parties can check whether or not i belongs to their sets.

Now we look at $x_i = y_i = 1$. We show that the protocol over a *random* choice of i should output i with the expected probability, that is, $1/|S \cap T|$. Note that the expected size of the intersection is

$$\mathbb{E}[|S \cap T|] = \mathbb{E}[\sum_{i=1}^{N} x_i y_i] = \sum_{i=1}^{N} \mathbb{E}[x_i y_i] = Npq \ ,$$

where we have used the linearity of expectation and independence of x_i and y_i. We proceed as follows.

$$\begin{aligned}
&\mathbb{E}_i[\Pr[\Pi^i_{1,1}(X,Y)=i]] = \\
&\qquad = \frac{1}{N}\sum \Pr[\Pi(X,Y)=i | x_i=y_i=1] \\
&\qquad = \frac{1}{Npq}\sum \Pr[\Pi(X,Y)=i \wedge x_i=y_i=1] \\
&\qquad = \frac{1}{Npq}\sum_i \sum_{(x,y):\ x_i=y_i=1\wedge\pi(x,y)=i} \Pr[(X,Y)=(x,y)] \\
&\qquad = \frac{1}{Npq}\sum_{(x,y):\ \pi(x,y)\text{ correct and }x\cap y\neq\emptyset} \Pr[(X,Y)=(x,y)] \\
&\qquad = \frac{1}{Npq}\Big(\sum_{(x,y)}\Pr[(x,y)] - \sum_{x\cap y=\emptyset}\Pr[(x,y)] - \sum_{\pi(x,y)\text{ fails}}\Pr[(x,y)]\Big) \\
&\qquad \geq \frac{1-\Pr[\mathrm{DISJ}(S,T)=1]-\varepsilon}{Npq} = \frac{p_0-\varepsilon}{Npq}\ .
\end{aligned}$$

Thus we get that

$$\mathbb{E}_i[\Delta_{\mathsf{TV}}(\Pi^i_{0,0},\Pi^i_{1,1})] \geq (p_0-\varepsilon)/(Npq)\ ,$$

and overall we obtain

$$\mathsf{D}^{\varepsilon}_{\mu(p,q)}(\mathrm{INT}) \geq \frac{Np}{8}\cdot\Big(\frac{p_0-\varepsilon}{Npq}\Big)^2\ ,$$

as required. □

Letting $p = 1/N^\alpha$ and $q = 1/N^\beta$ with $\alpha \geq \beta$ (since we assumed $p \leq q$), for a non-trivial lower bound—that is an exponentially large right-hand side in the displayed equation above—we would need to have that $\alpha + 2\beta > 1$. We also require that $1-\alpha-\beta > 0$ so that the expected intersection size Npq is exponentially large, in which case $p_0 \approx 1$ and set-disjointness is *easy*. These inequalities lead to the feasibility region shown in Figure 3. We have included the symmetric region for $\alpha \leq \beta$.

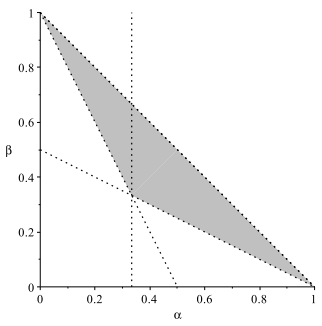

Fig. 3: Region where set intersection is hard with $p = 1/N^\alpha$ and $q = 1/N^\beta$.

In this work, we will rely on set-disjointness and set-intersection problems, as well as the following *multi-set* extensions of them. These problems are additionally parameterized by the number of sets. Here Alice holds M_1 sets $S_i \sim \mathrm{Bin}(N,p)$ for $i \in [M_1]$ and Bob holds M_2 sets $T_j \sim \mathrm{Bin}(N,q)$ for $j \in [M_2]$. Their goal is to solve the following problems.

1. Find (i, x) such that $x \in S_i \cap T_i$, or return $\perp$ if all the intersections are empty. We call this the (M_1, M_2)-INT problem, a natural multi-instance version of INT. A decisional variant would ask for an index i and a decision bit indicating if $S_i \cap T_i = \emptyset$. When $M_1 = M_2 = 1$, these problems are the usual INT and DISJ problems.
2. Find (i, j, x, x') with $x \neq x'$ such that $x, x' \in S_i \cap T_j$, or return $\perp$ if no such tuple exists. We call this the (M_1, M_2)-2INT problem. When $M_1 = M_2 = 1$ this problem is at least as hard as the INT problem since finding two distinct elements in the intersection is harder than finding one.

Remark. Intuitively, the INT problem is a harder task than (M_1, M_2)-2INT. One can solve the (M_1, M_2)-2INT problem using a protocol for INT as follows. Alice chooses a random point x in one of its sets S_i and sends it to Bob. Bob will then search through his sets to find a set T_j such that $x \in T_j$. With high probability such a set exists if the number of sets and/or the probability parameters are large enough. Alice and Bob will then run the protocol for INT on sets S_i and T_j to find an $x' \in S_i \cap T_j$. This element will be different from x with good probability (again under appropriate choices of parameters). Indeed, this is simply the communication complexity way of saying "collision-resistance implies one-wayness." However, we are interested in a reduction in the converse direction (as we already have lower bounds for INT). This seems hard as from a cryptographic point of view, as a classical impossibility by Simon [35] shows that collision resistance cannot be based on one-way functions (or even permutations) in a black-box way. Despite this, it is conceivable that direct information-theoretic analyses (similar to those for set-disjointness and set-intersection) can lead to non-trivial lower bounds. We leave proving hardness for this "collision resistance" analogue of set-intersection as an interesting open problem for future work.[9]

5 The Concatenation Combiner

In this section we study the security of the concatenation combiner

$$\mathsf{C}^{\mathsf{H}_1,\mathsf{H}_2}(x) := \mathsf{H}_1(x)|\mathsf{H}_2(x)$$

in the 2-BRO model, where $\mathsf{H}_1 \in \mathrm{Fun}[n, n + s_1]$ and $\mathsf{H}_2 \in \mathrm{Fun}[n, n + s_2]$. We will prove one-way security, pseudorandomness, and collision resistance for this construction. Our results will rely on the hardness of set-intersection and set-disjointness for the first two properties, and the presumed hardness of finding two elements in the intersection given multiple instances.

[9] We note that our setting is different to direct sum/product theorems where the focus is on hardness amplification. One shows, for example, that computing n independent copies of a function requires n times the communication for one copy for product distributions [3].

5.1 One-way security

In Appendix D we show that when H_1 or H_2 is (approximately) length preserving or somewhat expanding the concatenation combiner is not (strongly) one-way in the 2-BRO model. In both cases preimage sets will be only polynomially large and can be communicated. Accordingly, only when both hash functions are (somewhat) compressing we can achieve one-way security.

To this end, we first give a direct reduction from random preimage resistance (rPre, as defined in Figure 2) to set-intersection. By Lemma 2 we know that any (weak) rPre-secure function is also a (weak) OW-secure function. In particular, for the highly compressing setting where $s_1, s_2 \leq -n/2 - 4$ we show *strong* one-way security. For settings where the parameters only enable weak security according to the set-intersection theorem, we can apply hardness amplification [17] to get a strongly one-way function.

In our reductions to communication complexity protocols throughout the paper, we make the following simplifying assumptions. 1) The adversary is deterministic; 2) It does not query H_i at all and instead computes hash values via the BD_i oracles; 3) It queries BD_i with functions that have 1-bit outputs; and 4) It starts with a query to BD_1.

We are now ready to prove our first cryptographic hardness result.

Theorem 4. *Let* $\mathsf{H}_1 \in \text{Fun}[n, n+s_1]$ *and* $\mathsf{H}_2 \in \text{Fun}[n, n+s_2]$ *and* $\mathsf{C}^{\mathsf{H}_1,\mathsf{H}_2}(x) := \mathsf{H}_1(x)|\mathsf{H}_2(x)$. *Then for any adversary* $\mathcal{A}$ *against the* rPre *security of* $\mathsf{C}^{\mathsf{H}_1,\mathsf{H}_2}$ *in the* 2*-BRO model there is a* 2*-party protocol* π *against set-intersection with* $\mu := \mu(p,q)$ *where* $p := 1/2^{n+s_1}$ *and* $q := 1/2^{n+s_2}$ *and such that*

$$\mathsf{Adv}^{\text{rpre}}_{\mathsf{C}^{\mathsf{H}_1,\mathsf{H}_2}}(\mathcal{A}) \leq \mathsf{Adv}^{\text{int}}_{\mu}(\pi) \qquad \textit{and} \qquad \mathsf{D}_\mu(\pi) \leq \mathsf{Q}(\mathcal{A}) + 3n + s_1 + s_2 \ .$$

Proof. Let $\mathcal{A}$ be an adversary against the rPre security of $\mathsf{C}^{\mathsf{H}_1,\mathsf{H}_2}$ in the 2-BRO model for H_1 and H_2. Adversary $\mathcal{A}$ is given a random point $y := y_1|y_2 \in \{0,1\}^{2n+s_1+s_2}$ and needs to either find an x such that $\mathsf{H}_1(x)|\mathsf{H}_2(x) = y_1|y_2$ or say that no such x exists. Let

$$S_1 := \mathsf{H}_1^-(y_1) \qquad \text{and} \qquad S_2 := \mathsf{H}_2^-(y_2) \ .$$

Hence $\mathcal{A}$ outputs an $x \in S_1 \cap S_2$ as long as $S_1 \cap S_2 \neq \emptyset$. We note that these sets are Bernoulli. Indeed, for each x we have that $\Pr[x \in S_1] = 1/2^{n+s_1}$ and $\Pr[x \in S_2] = 1/2^{n+s_2}$, and these events are independent for different values of x.

We use $\mathcal{A}$ to build a 2-party protocol for set-intersection over a product distribution $\mu := \mu(p,q)$ with $p := 1/2^{n+s_1}$ and $q := 1/2^{n+s_2}$ as follows. Alice holds a set $S_1 \subseteq \{0,1\}^n$ and Bob holds a set $S_2 \subseteq \{0,1\}^n$ distributed according to μ. Alice (resp., Bob) samples hash function H_1 (resp., H_2) as follows. Alice picks a random $y_1 \in \{0,1\}^{n+s_1}$ and Bob picks a random $y_2 \in \{0,1\}^{n+s_2}$. Alice defines H_1 to map all points in S_1 to y_1. She maps $x \in \{0,1\}^n \setminus S_1$ to random points in $\{0,1\}^{n+s_1} \setminus \{y_1\}$. Similarly Bob defines H_2 to map all points $x \in S_2$ to y_2 and $x \in \{0,1\}^n \setminus S_1$ to random points in $\{0,1\}^{n+s_2} \setminus \{y_2\}$. As a result, Alice knows the full function table of H_1 and similarly Bob knows the full function table of H_2.

Alice and Bob now run two copies of $\mathcal{A}$ in tandem as follows, where the state values st_A and st_B are initially set to $y_1|y_2$ (with only $2n + s_1 + s_2$ bits of communication).

Alice: It resumes/starts $\mathcal{A}(st_A)$. It terminates if it receives a final guess x from Bob. It answers all pending BD_2 queries—there are none to start with—using the values just received from Bob. It answers all BD_1 queries using the function table of H_1 until $\mathcal{A}$ queries BD_2 or terminates. If $\mathcal{A}$ terminates with a final guess x, it forwards x to Bob and terminates. Else it saves the current state st_A of $\mathcal{A}$ locally and forwards all BD_1 answers that it has provided to $\mathcal{A}$ since the last resumption to Bob. It hands the execution over to Bob.

Bob: It resumes $\mathcal{A}(st_B)$. It terminates if it receives a final guess x from Alice. It answers all pending BD_1 queries using the values received from Alice. It answers all BD_2 queries using the function table of H_2 until $\mathcal{A}$ queries BD_1 or terminates. If $\mathcal{A}$ terminates with a final guess x, it forwards x to Alice and terminates. Else it saves the current state st_B of $\mathcal{A}$ locally and forwards all BD_2 answers that it has provided to $\mathcal{A}$ since the last resumption to Alice. It hands the execution over to Alice.

We claim that Alice and Bob run $\mathcal{A}$ in an environment that is identical to the rPre game in the 2-BRO model. The hash functions H_1 and H_2 sampled by Alice and Bob are uniformly distributed. To see this note that for any (x, y) the probability that $\mathsf{H}_1(x) = y$ is $1/|\{0,1\}^{n+s_1}|$ (and similarly for H_2). Furthermore, this event is independent of the hash values that are set for all other values $x' \neq x$. Thus, Alice and Bob faithfully run $\mathcal{A}$ in the environment that it expects by answering its backdoor queries using their knowledge of the full tables of the two functions.

Whenever $\mathcal{A}$ succeeds in breaking the rPre security of $\mathsf{C}^{\mathsf{H}_1,\mathsf{H}_2}$, the protocol above computes an $x \in S_1 \cap S_2$ or says that no such x exists. In either case, the protocol solves the set-intersection problem. Thus the correctness of this protocol is at least the advantage of the adversary $\mathcal{A}$.

This execution of $\mathcal{A}$ by Alice and Bob ensures that oracle *queries* do not affect the communication cost of Alice and Bob. It is only their answers (plus the final x) that affects the communication cost, since the queried functions f are locally computed and only their answers are communicated. If $\mathcal{A}$ makes $\mathsf{Q}(\mathcal{A})$ queries to BD_1 and BD_2 in total and each query has a 1-bit output, the total communication complexity of the protocol is $\mathsf{Q}(\mathcal{A})$ plus those bits needed to communicate y_1 and y_2 and the final guess x. □

We now check that the parameters for hash functions can be set such that their concatenation is a one-way function.

Corollary 1. *For* $\mathsf{H}_1, \mathsf{H}_2 \in \text{Fun}[n, (1-\epsilon)n/2]$ *with* $0 < \epsilon < 1/3$ *the concatenation combiner is a strongly one-way compressing function in* $\text{Fun}[n, (1-\epsilon)n]$.

Proof. The feasible region in Figure 3 for $\alpha = \beta$ consists of $1/3 < \alpha < 1/2$. In our setting $\alpha = \beta = (1-\epsilon)/2$, which means concatenation is strongly rPre secure when $0 < \epsilon < 1/3$. Since the combined function is compressing (where

$\gamma = 1/(1-\epsilon) > 1$), the image-uniformity bound is negligible and also $\Pr[y \in \mathsf{Img}(\mathsf{C}^{\mathsf{H}_i})]$ in Lemma 2 is overwhelming. Using these bounds and Lemma 2 we get that strong rPre security implies strong OW security. □

We conjecture that concatenation is strongly one-way even for $1/3 \le \epsilon < 1$. The intuition is that in the one-way game a point is "planted" in a large intersection, which seems hard to discover without essentially communicating the entire intersection. Tighter lower bounds for set-intersection can be used to establish this.

5.2 PRG security

We now consider the PRG security of the concatenation combiner. Our reduction in Theorem 4 from rPre to set-intersection can be easily adapted to the decisional setting. That is, we can show that a decisional variant of rPre can be reduced to the set-disjointness problem. The decisional variant of rPre asks the adversary to decide whether or not a random co-domain point $y_1|y_2$ has a preimage. This is exactly the oblivious PRG (oPRG) notion that we defined in Section 2. We get the following result.

Theorem 5. *Let* $\mathsf{H}_1 \in \mathrm{Fun}[n, n+s_1]$ *and* $\mathsf{H}_2 \in \mathrm{Fun}[n, n+s_2]$ *and* $\mathsf{C}^{\mathsf{H}_1,\mathsf{H}_2}(x) := \mathsf{H}_1(x)|\mathsf{H}_2(x)$. *Then for any adversary* $\mathcal{A}$ *against the oblivious PRG security of* $\mathsf{C}^{\mathsf{H}_1,\mathsf{H}_2}$ *in the* 2*-BRO model there is a* 2*-party protocol* π *against set-disjointness with* $\mu := \mu(p,q)$ *where* $p := 1/2^{n+s_1}$ *and* $q := 1/2^{n+s_2}$ *and such that*

$$\mathsf{Adv}^{\mathrm{oprg}}_{\mathsf{C}^{\mathsf{H}_1,\mathsf{H}_2}}(\mathcal{A}) \le \mathsf{Adv}^{\mathrm{disj}}_{\mu}(\pi) \qquad \textit{and} \qquad \mathsf{D}_{\mu}(\pi) \le \mathsf{Q}(\mathcal{A}) + 2n + s_1 + s_2 + 1 \ .$$

We next check if concrete parameters can be set to obtain an expanding PRG.

Corollary 2. *For* $s_1 = -n/2 + 1$ *and* $s_2 = -n/2$, *the concatenation combiner gives a weak* PRG *in* $\mathrm{Fun}[n, n+1]$.

Proof. The theorem gives a reduction to set-disjointness with parameters $p = 1/2^{n/2+1}$ and $q = 1/2^{n/2}$. For large n we get, $\delta = 2$, $p_0 = 1 - e^{-1/2}$ and $(\delta-1)p_0/(4+\delta) < 0.0656$, which means we can set $\varepsilon = 0.065$. By set-disjointness lower bound, this means any adversary with advantage at least 0.935 must place at least $\mathcal{O}(2^{n/2})$ queries in total to its oracles.

By Lemma 1 we have that $\mathsf{Adv}^{\mathrm{iu}}_{\mathsf{C}^{\mathsf{H}_i}}(\mathcal{B}) \le e^{-C} \cdot \left(C/(1-e^{-C}) - 1\right)$. In our case $C = 1/2 < 1$, and the right hand side above is upper bounded by ≤ 0.165. (We have removed the negligible terms and instead approximated the constants by slightly larger values.)

In Lemma 3 in order to meet the bound $\mathsf{Adv}^{\mathrm{oprg}}_{\mathsf{C}^{\mathsf{H}_i}}(\mathcal{C}) < (2 - \alpha - \mathsf{Adv}^{\mathrm{iu}}_{\mathsf{C}^{\mathsf{H}_i}}(\mathcal{B})) \cdot \alpha/(1-\alpha)$, we would need $0.935 \le (2-\alpha-0.165)\cdot\alpha/(1-\alpha)$. After some algebra this gives $\alpha \ge 0.39343$. With $m = n + s$, we need to have $1 - e^{-2^{-s}} \ge 0.39343$, which means $s \le 1.00018$. Thus we can set $s = 1$ (which also satisfies $s \ge 0.53$ as required in the lemma). □

We can obtain a strong PRG by amplification. However, we need an amplifier that woks on PRGs with (very) *small* stretch. Such a construction is given by Maurer and Tessaro [28]. In their so-called Concatenate-and-Extract (CaE) construction one sets

$$\mathsf{PRG}(r, x_1, \ldots, x_m) := r | \mathsf{Ext}\big(r, \mathsf{C}^{\mathsf{H}_1,\mathsf{H}_2}(x_1) | \cdots | \mathsf{C}^{\mathsf{H}_1,\mathsf{H}_2}(x_m)\big) \ ,$$

where Ext is a sufficiently good randomness extractor, for instance a two-universal hash function. We refer to the original work for concrete parameters. It is safe to assume the extractor is backdoor-free, since it is an information-theoretic object and relatively easy to implement.

5.3 Collision resistance

The classical result of Simon [35] shows that collision-resistance relies on qualitatively stronger assumptions than one-way functions. In the theorem below we prove collision resistance based on the hardness of the multi-instance 2INT problem as defined in Section 4. As discussed in the final remark of that section, we do not expect that a reduction to the INT problem exists.

Theorem 6. *Let* $\mathsf{H}_1 \in \mathrm{Fun}[n, n+s_1]$ *and* $\mathsf{H}_2 \in \mathrm{Fun}[n, n+s_2]$ *and* $\mathsf{C}^{\mathsf{H}_1,\mathsf{H}_2}(x) := \mathsf{H}_1(x) | \mathsf{H}_2(x)$*. Then for any adversary* $\mathcal{A}$ *against the collision resistance of* $\mathsf{C}^{\mathsf{H}_1,\mathsf{H}_2}$ *in the 2-BRO model there is a 2-party protocol* π' *against multi-instance two-element set-intersection problem over* $\mu' := \mu(p', q')$ *with* $p' := 2n \ln 2 / 2^{n+s_1}$ *and* $q' := 2n \ln 2 / 2^{n+s_2}$ *and where Alice holds* $M_1 := 2^{n+s_1}$ *sets and Bob holds* $M_2 := 2^{n+s_2}$ *sets such that*

$$\mathsf{Adv}^{\mathrm{cr}}_{\mathsf{C}^{\mathsf{H}_1,\mathsf{H}_2}}(\mathcal{A}) \le \mathsf{Adv}^{\text{mi-2int}}_{\mu'}(\pi') + 2 \cdot 2^{-n} \qquad \textit{and} \qquad \mathsf{D}_{\mu'}(\pi') \le \mathsf{Q}(\mathcal{A}) + 4n + s_1 + s_2 \ .$$

Proof. We follow an overall strategy that is similar to one for the rPre reduction. For each $i \in \{0,1\}^{n+s_1}$, Alice sets $\mathsf{H}_1^-(i) := S_i$ and for each $j \in \{0,1\}^{n+s_2}$ Bob sets $\mathsf{H}_2^-(j) := T_j$ and they simulate the two hash functions. However, this leads to a problem: S_i are not necessarily disjoint and furthermore their union does not cover the entire domain $\{0,1\}^n$. (The same is true for T_j.) Put differently, the distributions of sets formed by hash preimages of co-domain points do not match independently chosen sets from a Bernoulli distribution.

We treat this problem in two step. The first step is a direct reduction to a "partitioned" modification of the multi-instance set-intersection. In this partitioned problem Alice gets sets $S_i := \mathsf{H}_1^-(i)$ for $i \in \{0,1\}^{n+s_1}$ and a random oracle $\mathsf{H}_1 \in \mathrm{Fun}[n, n+s_1]$. Similarly, Bob gets sets $T_j := \mathsf{H}_2^-(j)$ for $j \in \{0,1\}^{n+s_2}$ and an independent random oracle $\mathsf{H}_2 \in \mathrm{Fun}[n, n+s_2]$. Their goal is to find a tuple (i, j, x, x') with $x \neq x'$ such that $x, x' \in S_i \cap T_j$. Thus, these sets exactly correspond to hash preimages as needed in the reduction above, and a solution would translate to a collision for the combined hash function.

We then show that hardness of the (standard) multi-instance two-element set-intersection problem implies the hardness of the partitioned problem with an increase in the Bernoulli *parameter*.[10] □

[10] Another strategy would be to change the *number* of sets involved. But this runs into a problem as this number must match the size of the co-domain of the hash function.

Algo. $\mathsf{ReDist}(S_1, \ldots, S_M)$

for $x \in \{0,1\}^n$ **do**
  $A_x := \{i \in [M] : x \in S_i\}$
  $i_x \twoheadleftarrow A_x$
  for $j \in [M] \wedge j \neq i_x$ **do**
    $\widetilde{S}_j \leftarrow \widetilde{S}_j \setminus \{x\}$
return $(\widetilde{S}_1, \ldots, \widetilde{S}_M)$

Fig. 4: Redistribution of elements to form a partition.

Lemma 6 (Partitioned $\implies$ Independent). *For any two-party protocol π against the partitioned multi-instance set-intersection problem there is a two-party protocol π' against multi-instance set-intersection problem such that*

$$\mathsf{Adv}^{\text{mi-2int}}_{\mu'}(\pi') \geq \mathsf{Adv}^{\text{part-2int}}_{\mu}(\pi) - 2 \cdot 2^{-n} \qquad \textit{and} \qquad \mathsf{D}_{\mu'}(\pi') \leq \mathsf{D}_{\mu}(\pi) \ .$$

Here $\mu := \mu(p,q)$ is the distribution induced by hash preimages and $\mu' := \mu'(p',q')$ is a product Bernoulli with $p' := 2n \ln 2 \cdot p$ and $q' := 2n \ln 2 \cdot q$.

Proof. To focus on the core ideas, we simplify and let $M_1 = M_2 = M = 2^{n+s}$ and $p = q = 1/2^{n+s}$. Suppose we have sets S_i and T_j for $i = 1, \ldots, M$ and $j = 1, \ldots, M$ as an instance for the multi-instance intersection. Let $p' = q' = 2n \ln 2 \cdot p$. Then

$$\begin{aligned} \Pr[\exists x \in \{0,1\}^n \ \forall i \in [M] : x \notin S_i] &\leq 2^n \Pr[\forall i \in [M] : x \notin S_i] \\ &\leq 2^n (1-p')^{1/p} \leq 2^n e^{-2n \ln 2} = 2^{-n} \ . \end{aligned}$$

Thus with these parameters the sets S_i (and similarly T_j) will cover the full domain, that is $\bigcup_{i=1}^{M} S_i = \{0,1\}^n$.

Note that with these parameters any two sets S_i and T_j will intersect with overwhelming probability. However, finding an element in the intersection may still be hard; see conjecture below.

Our next step it to redistribute the elements among the sets so that they form partitions. We do this via the algorithm ReDist shown in Figure 4. ReDist iterates through elements x in the domain and leaves x in exactly one of the sets. (By the above covering property such a set always exits.)

This procedure will be applied to S_i (resp., T_j) to produce non-overlapping sets $\tilde{S}_i$ (resp. $\tilde{T}_j$). Furthermore, we always have that $\tilde{S}_i \subseteq S_i$ and $\tilde{T}_i \subseteq T_i$, since elements are only deleted from the sets and never added to them. Thus $\tilde{S}_i \cap \tilde{T}_j \subseteq S_i \cap T_j$ as well, and this means that any solution with respect to the tweaked sets will also be a valid solution for the original (Bernoulli) sets.

We still need to show that the distribution of the tweaked sets is identical to that given by hash preimages under a random oracle. Let $E_{x,i}$ be the event

that $x \in \widetilde{S}_i$. Since the algorithm does not treat any of the i's in a special way, we claim that $\Pr[E_{x,i}]$ is independent of i. Indeed for any i, j we have

$$\Pr[E_{x,i}] = \Pr[x \in S_i] \Pr[i_x = i | x \in S_i] = \Pr[x \in S_j] \Pr[i_x = j | x \in S_j] = \Pr[E_{x,j}].$$

This is because $\Pr[x \in S_i] = \Pr[x \in S_j]$ and $\Pr[i_x = i | x \in S_i] = \Pr[i_x = j | x \in S_j]$. If we call this common probability e_x, since x is guaranteed belongs to one of the M sets, we have that $\sum_{i \in [M]} e_x = 1$. Thus $e_x = 1/M = \Pr[\mathsf{H}_1(x) = i]$. Note that the algorithm assigns different values of x independently of all other values already assigned, we get that the event $\mathsf{H}_1(x) = i$ is independent for different x.

Finally, solutions with respect to the tweaked sets always exist when $s_1 + s_2 < 0$. This is because the problem is equivalent to finding collisions for a function $\mathsf{H}_1(x) | \mathsf{H}_2(x)$ that is compressing, which necessarily exist. □

The birthday attack gives a $2^{\min(n+s_1, n+s_2)/2}$ upper bound on the security of the combined hash function. Balancing the digest lengths with $s_1 = s_2 = n/2$, leads to a maximum collision security of at most $2^{n/4}$. Proving a lower bound, on the other hand, remains an interesting open problem. We formulate a conjecture towards proving this next.

Conjecture 1. The multi-instance two-element set-intersection problem over Bernoulli sets in a universe of size N with $p = q = 1/\sqrt{N}$ and $\sqrt{N}$ sets for each party has communication complexity

$$\mathsf{D}^{\varepsilon}_{\mu(p,q)}((\sqrt{N}, \sqrt{N})\text{-2INT}) \geq \tilde{\Omega}(N^{1/4}),$$

for a sufficiently small protocol error ε and where $\tilde{\Omega}$ hides logarithmic factors.

We note that a lower bound for protocols with a sufficiently small error would be sufficient for feasibility results as collision resistance can also be amplified in a black-box way [7].

6 The Cascade Combiner

We now look at the security of the cascade combiner

$$\mathsf{C}^{\mathsf{H}_1, \mathsf{H}_2}(x) := \mathsf{H}_2(\mathsf{H}_1(x))$$

in the 2-BRO model, where $\mathsf{H}_1 \in \mathrm{Fun}[n, n + s_1]$ and $\mathsf{H}_2 \in \mathrm{Fun}[n + s_1, n + s_1 + s_2]$. We will prove one-way security and pseudorandomness based on set-intersection and set-disjointness respectively, and collision resistance based on a variant finding two intersecting points given multiple instances for one party and a single set for the other.

6.1 One-way security

Similarly to the concatenation combiner, we can reduce the random preimage resistance (rPre) security of the cascade combiner to set-intersection.

Theorem 7. *Let* $\mathsf{H}_1 \in \mathrm{Fun}[n, n+s_1]$ *and* $\mathsf{H}_2 \in \mathrm{Fun}[n+s_1, n+s_1+s_2]$ *and* $\mathsf{C}^{\mathsf{H}_1,\mathsf{H}_2}(x) := \mathsf{H}_2(\mathsf{H}_1(x))$. *Then for large enough* n *and any adversary* $\mathcal{A}$ *against the* rPre *security of* $\mathsf{C}^{\mathsf{H}_1,\mathsf{H}_2}$ *in the* 2*-BRO model there is a* 2*-party protocol* π *against set-intersection with* $\mu := \mu(p,q)$ *where* $p := 1/2^{s_1}$ *and* $q := 1/2^{n+s_1+s_2}$ *and such that*

$$\mathsf{Adv}^{\mathrm{rpre}}_{\mathsf{C}^{\mathsf{H}_1,\mathsf{H}_2}}(\mathcal{A}) \le \mathsf{Adv}^{\mathrm{int}}_{\mu}(\pi) + \sqrt{n}2^{-n/2}(1+2^{-s_2-s_1})$$
$$\textit{and} \quad \mathsf{D}_{\mu}(\pi) \le \mathsf{Q}(\mathcal{A}) + 3n + s_1 + s_2.$$

Proof. We follow a strategy similar to the reductions in Section 5. Given a random $y^* \in \{0,1\}^{n+s_1+s_2}$ the task of the adversary $\mathcal{A}$ against the rPre security of $\mathsf{C}^{\mathsf{H}_1,\mathsf{H}_2}$ is to a find a z such that $\mathsf{C}^{\mathsf{H}_1,\mathsf{H}_2}(z) = y^*$. With such a z, one can then also compute $x := \mathsf{H}_1(z)$ and conclude that $x \in I \cap T$ where

$$I := \mathsf{H}_1(\{0,1\}^n) \qquad \text{and} \qquad T := \mathsf{H}_2^-(y^*)$$

with $I, T \subseteq \{0,1\}^{n+s_1}$. The set T is Bernoulli with parameter $\Pr[y \in T] = 1/2^{n+s_1+s_2}$. Although set I appears to be Bernoulli,

$$\Pr[x \in I] = 1 - \Pr[\forall z : \mathsf{H}_1(z) \neq x] = 1 - (1 - 1/2^{n+s_1})^{2^n}$$

it is not, since these probabilities are not independent for different values of x.

Our strategy to deal with this, and ultimately construct a protocol π for solving set-intersection, is to start with a Bernoulli set S (Alice's input), and program H_1 on all $x \in \{0,1\}^n$ to values y that will be taken from S, but are also set to *collide* with the "right" probability. This will ensure that the image of H_1 contains most of S and is also distributed as the image of a random oracle.

We proceed as follows. Initially the set of assigned domain points X and assigned co-domain points Y are empty. We then iterate through $x \in \{0,1\}^n$ in a random order. A bit b at each iteration decides if the hash value y for x should collide with a previously assigned value or not. If so, we sample y from the set of already assigned values Y. Otherwise, y should be a non-colliding value and we sample it from S if S is non-empty (and remove y from S), or otherwise we sample it outside the already assigned points Y. The pseudo-code for this algorithm, which we call $\mathsf{HashSam}$, is shown in Figure 5.

Setting $m := n+s_1$, we now need to check that (1) the returned hash function H_1 is distributed as a random oracle $\{0,1\}^n \to \{0,1\}^m$ when S is Bernoulli with parameter $p = 1/2^{s_1}$, and (2) if $x \in \mathsf{H}_1(\{0,1\}^n) \cap \mathsf{H}_2^-(y^*)$, then we also have that $x \in S \cap T$ with good probability.

We first prove (1). The intuition is that the algorithm treats all inputs and outputs in a uniform way, and hence no particular values are special. Formally, let x^* and y^* be any fixed values. We show that $\Pr[\mathsf{H}_1(x^*) = y^*] = 1/2^m$, even

Algo. $\mathsf{HashSam}(S)$

$X \leftarrow \emptyset;\ Y \leftarrow \emptyset$
for $i = 1, \ldots, 2^n$ **do**
  $x \twoheadleftarrow \{0,1\}^n \setminus X; X \leftarrow X \cup \{x\}$
  $b \twoheadleftarrow \mathrm{Ber}(|Y|/2^m)$
  if $b = 1$ **then** $y \twoheadleftarrow Y$
  if $b = 0 \wedge S = \emptyset$ **then**
    $y \twoheadleftarrow \{0,1\}^m \setminus Y;\ Y \twoheadleftarrow Y \cup \{y\}$
  if $b = 0 \wedge S \neq \emptyset$ **then**
    $y \twoheadleftarrow S;\ Y \leftarrow Y \cup \{y\};\ S \leftarrow S \setminus \{y\}$
  $\mathsf{H}_1 \leftarrow \mathsf{H}_1 : [x \mapsto y]$
return H_1

Fig. 5: Hash sampler centered around a Bernoulli set S.

given the previously assigned values. We use a subscript i to denote the values of various variables in the i-th iteration. Looking at different execution branches of the algorithm we can calculate $\Pr[y_i = y^* | x_i = x^*, Y_i, X_i]$ as

$$\Pr[b_i = 1]\Pr[y^* \in Y_i]\frac{1}{|Y_i|} + \Pr[b_i = 0]\Big(\Pr[S_i = \emptyset]\Pr[y^* \notin Y_i]\frac{1}{2^m - |Y_i|} + \\ + \Pr[S_i \neq \emptyset]\Pr[y^* \in S_i]\frac{1}{|S_i|}\Big) .$$

Letting $\theta_i := \Pr[S_i = \emptyset]$ we can simplify to

$$\frac{|Y_i|}{2^m}\frac{|Y_i|}{2^m}\frac{1}{|Y_i|} + \Big(1 - \frac{|Y_i|}{2^m}\Big)\Big(\theta_i\Big(1 - \frac{|Y_i|}{2^m}\Big)\frac{1}{2^m - |Y_i|} + \Big(1 - \theta_i\Big)\frac{|S_i|}{2^m}\frac{1}{|S_i|}\Big) = \frac{1}{2^m} .$$

Note we have used the fact that S_i is a Bernoulli set in $\Pr[y^* \in S_i] = \frac{|S_i|}{2^m}$. Hence

$$\Pr[\mathsf{H}_1(x^*) = y^* | Y_i, X_i] = \sum_{i=1}^{2^n} \Pr[y_i = y^* | x_i = x^*, Y_i, X_i]\Pr[x_i = x^*] = \frac{1}{2^m} .$$

Therefore the probability of sampling any given hash function is $(1/2^m)^{2^n}$, as required.

Let us now consider (2). When $I \subseteq S$, any solution with respect to I is also one with respect to S (that is, solutions are not lost). Hence we only look at the case $S \subseteq I$ and bound $|I \setminus S| = |I| - |S|$. Since $|I| \leq 2^n$ and $\mathbb{E}[|S|] = 2^{n+s_1}/2^{s_1} = 2^n$, we get that for any t

$$\Pr[|I| - |S| > t] \leq \Pr[2^n - |S| > t] = \Pr[\mathbb{E}[|S|] - |S| > t] .$$

Applying the Chernoff bounds we obtain

$$\Pr\left[\mathbb{E}[|S|] - |S| > t\,\mathbb{E}[|S|]\right] \leq e^{-\frac{t^2}{2+t}\,\mathbb{E}[|S|]}\ .$$

Setting $t := \sqrt{n/2^n}$, we get with overwhelming probability (of at least $1 - \sqrt{n} \cdot 2^{-n/2}$) that $|I \setminus S| \leq \sqrt{n}2^{n/2}$. Hence $T \cap (I \setminus S)$ will be non-empty with negligible probability $\sqrt{n}2^{-n/2-s_1-s_2}$, in which case if $x \in I \cap T \implies x \in S \cap T$. □

If $\mathsf{H}_1 \in \mathrm{Fun}[n, (2+\epsilon)n]$ and $\mathsf{H}_2 \in \mathrm{Fun}[(2+\epsilon)n, (1+\epsilon)n])$, we have a reduction to set-intersection with parameters $N = 2^{(2+\epsilon)n}$, $p = 1/2^{(1+\epsilon)n}$, and $q = 1/2^{(1+\epsilon)n}$. Thus with notation as in the description of the feasible region in Figure 3 we have that $\alpha = \beta = (1+\epsilon)/(2+\epsilon)$. As in Corollary 1 we would need that the point (α, β) lies in the feasible region for $1/3 < (1+\epsilon)/(2+\epsilon) < 1/2$, which means $-1/2 < \epsilon < 0$. Since the combined function is compressing (with $\gamma = 1/(1+\epsilon) > 1$) and $p_\circ \approx 1 - e^{-2^{-\epsilon \cdot n}}$ is negligible, the image uniformity bound is negligible and also $\Pr[y \in \mathsf{Img}(\mathsf{C}^{\mathsf{H}_i})]$ in Lemma 2 is overwhelming. Hence, similarly to Corollary 1 we get strong OW security.

6.2 PRG and CR security

We briefly outline how to treat the PRG security and collision resistance of cascade. We omit the proofs as the techniques and proof structures are similar to our other results above.

PRG SECURITY. We can prove an analogous result for the oblivious PRG security of the cascade construction. Its reduction is identical to that for rPre security given above, except that the underlying assumption is set-disjointness. Setting $s_1 = 2n$ (H_1 is length doubling) and $s_2 = -2n + 1$ (H_2 compresses by almost a factor of 3) leads to a reduction to an instance of set-intersection with parameters $N = 2^{3n}$, $p = 1/2^{2n}$, and $q = 1/2^{n+1}$. In this case $\delta = 2$ and $p_0 = 1 - e^{-1/2}$. With these parameters we can carry out an analysis similar to Corollary 2: We set the error $\varepsilon = 0.065$ which is smaller than $(\delta - 1)p_0/(4+\delta) < 0.0656$ as required in Theorem 2 for an exponential number of queries. The combined hash function maps n bits to $n+1$ bits and hence $C = 1/2$. Furthermore, $p_\circ$ is negligible as a function from n bits to $3n$ bits is injective with overwhelming probability. Thus we can apply Lemma 3 with $s = 1$ as in Corollary 2 to get a weak PRG.

COLLISION RESISTANCE. We can treat the collision resistance of cascade similarly. The difference is that in the reduction Alice will use the $\mathsf{HashSam}$ algorithm in Figure 5 to adapt a (single) Bernoulli set S that she holds to a hash image set I. On the other hand, Bob uses the ReDist algorithm in Figure 4 to redistribute elements in multiple Bernoulli sets that he holds so that they form a partition of the entire domain of H_2. The rest of the proof, which is included in Appendix E, proceeds similarly to Lemma 6. For setting parameters, observe that any collision for H_1 is necessarily a collision for $\mathsf{H}_2(\mathsf{H}_1(\cdot))$. Since collisions for H_1 can be easily found using BD_1, we need H_1 to be injective. For example, $s_1 = 2n$ (co-domain points are $3n$ bits) would lead to an injective H_1 with overwhelming probability.

7 The xor Combiner

In this section we study the security of the xor combiner

$$\mathsf{C}^{\mathsf{H}_1,\mathsf{H}_2}(x) := \mathsf{H}_1(x) \oplus \mathsf{H}_2(x)$$

in the 2-BRO model, where both $\mathsf{H}_1, \mathsf{H}_2 \in \mathrm{Fun}[n, n+s]$. We will prove one-way security based on the hardness of the set-intersection problem and briefly discuss PRG security and collision resistance.

7.1 One-way security

We prove rPre-security of the xor combiner similarly to the previous combiners. Although the communication complexity problem directly underlying our reduction is multi-instance, we can still relate it to a standard single-instance problem, for which lower bounds are known.

Theorem 8. *Let* $\mathsf{H}_1, \mathsf{H}_2 \in \mathrm{Fun}[n, n+s]$ *and* $\mathsf{C}^{\mathsf{H}_1,\mathsf{H}_2}(x) := \mathsf{H}_1(x) \oplus \mathsf{H}_2(x)$*. Then for any adversary* $\mathcal{A}$ *against the* rPre *security of* $\mathsf{C}^{\mathsf{H}_1,\mathsf{H}_2}$ *in the* 2*-BRO model there is a* 2*-party protocol* π *against set-intersection with* $\mu = \mu(p', q')$ *where* $p' = q' = 2n \ln 2 / 2^{n+s}$ *and such that*

$$\mathsf{Adv}^{\mathrm{rpre}}_{\mathsf{C}^{\mathsf{H}_1,\mathsf{H}_2}}(\mathcal{A}) \le \mathsf{Adv}^{\mathrm{int}}_{\mu}(\pi) + 2 \cdot 2^{-n} \qquad \textit{and} \qquad \mathsf{D}_{\mu}(\pi) \le \mathsf{Q}(\mathcal{A}) + 3n + 2s \ .$$

Proof. Similarly to Theorem 6 we present the proof in two steps: we identify the underlying communication complexity problem and then relate it to a more standard/natural one.

In the rPre game for the xor combiner an adversary $\mathcal{A}$ is given a random point $y^* \in \{0,1\}^{n+s}$ and its task is to find a point $x^* \in \{0,1\}^n$ such that $\mathsf{H}_1(x^*) \oplus \mathsf{H}_2(x^*) = y^*$. Such an adversary exists if and only if there is one that can output a pair (i, x^*) such that $\mathsf{H}_1(x^*) = i$ and $\mathsf{H}_2(x^*) = i \oplus y^*$. That is, this adversary finds an $x^* \in S_i \cap T_{i \oplus y^*}$, where $S_i := \mathsf{H}_1^-(i)$ and $T_{i \oplus y^*} := \mathsf{H}_2^-(i \oplus y^*)$.

We simplify further. Given y^*, if we define $\overline{\mathsf{H}}_2(x^*) := \mathsf{H}_2(x^*) \oplus y^*$, the problem becomes equivalent to finding an $x^* \in S_i \cap \overline{T}_i$ where S_i is as before and $\overline{T}_i := \overline{\mathsf{H}}_2^-(i)$. Thus the problem at hand is $(2^{n+s}, 2^{n+s})$-INT, with multiple instances and parameters $p = q = 1/2^{n+s}$.

We can relate this problem to the single-instance set-intersection as follows. Suppose we have an instance (S^*, T^*) for set-intersection with parameters $p' := q' := 2n \ln 2 \cdot p$ over subsets of $\{0,1\}^{(n+s)+n}$. We consider elements $x = x_1 | x_2 \in S^*$ where $x_1 \in \{0,1\}^{n+s}$ and $x_2 \in \{0,1\}^n$, and compute the probability that the x_2's cover $\{0,1\}^n$ as follows.

$$\begin{aligned} \Pr[\exists x_2 \in \{0,1\}^n \ \forall x_1 \in \{0,1\}^{n+s} : (x_1|x_2) \notin S^*] & \\ \le 2^n \Pr[\forall x_1 \in \{0,1\}^{n+s} : (x_1|x_2) \notin S^*] & \\ \le 2^n (1-p')^{2^{n+s}} \le 2^n e^{-2n \ln 2} = 2^{-n} & \end{aligned}$$

Thus with parameter p' the x_2 in S^* will cover $\{0,1\}^n$ with overwhelming probability. A similar result holds for T^*. Thus if we define $S_{x_1} := \{x_2\}$ and $\overline{T}_{x_1} := \{x_2\}$ we will have with overwhelming probability that

$$\bigcup_{x_1 \in \{0,1\}^{n+s}} S_{x_1} = \{0,1\}^n \qquad \text{and} \qquad \bigcup_{x_1 \in \{0,1\}^{n+s}} \overline{T}_{x_1} = \{0,1\}^n \ .$$

Our next step it to redistribute the elements among the sets so that form partitions. As in Lemma 6, we can do this using the ReDist algorithm of Figure 4. As shown in the proof of Lemma 6, this procedure will produce a random oracle and hence the sets S_i and $\overline{T}_j$ will be correctly distributed as preimages under two random oracles. Since elements are only deleted from the original (Bernoulli) sets, any solution to rPre game is also one for the original sets. □

If $\mathsf{H}_1, \mathsf{H}_2 \in \mathrm{Fun}[n, (1+\epsilon)n]$, we have a reduction to the single-instance set intersection with $p' = q' = 2n \ln 2 / 2^{(1+\epsilon)n}$ and a universe of size $N = 2^{n+(1+\epsilon)n}$. We can write p' (and q') as $1/N^\alpha$ with $\alpha = \frac{(1+\epsilon)n - \log(2n \ln 2)}{(2+\epsilon)n}$. The feasible region in Figure 3 for $\alpha = \beta$ consists of $1/3 < \alpha < 1/2$. These inequalities for $n \geq 128$ translate to $-\frac{13}{32} + \frac{3}{256}\log(\ln 2) \approx -0.42 < \epsilon < 0$. For such parameters we thus get strong rPre security. Furthermore, since $\epsilon < 0$ the functions are (highly) compressing, which means the image uniformity bound is negligible. With an analysis similar to Corollary 1 we get strong one-way security.

For oPRG security, a similar reduction to set-disjointness can be given. With $\mathsf{H}_1, \mathsf{H}_2 \in \mathrm{Fun}[n, n+s]$, we have $N = 2^{2n+s}$ and $p' = q' = 2n \ln 2 / 2^{n+s}$ in the reduction and hence $\delta = 1/(Npq) = 2^s/(2n \ln 2)^2$. We set $s = 2\log n + 2$, in which case we have that $\delta = 1/(\ln 2)^2 > 1$, and the set-disjointness lower bound is non-trivial (and is at least $2^{n - \mathcal{O}(\log n)}$) when $\varepsilon < 0.067$. This establishes the oPRG security of xor for $s = 2\log n + 2$. However, an attempt to show weak PRG security (as for the other combiners) fails. The term corresponding to image uniformity bound in Lemma 1 when $s = 2\log n + 2$ (i.e., when $C := 1/2^s = 1/(4n^2)$) is less than 0.001 for $n \geq 12$ (and hence small enough). But in order to upper bound the PRG advantage via Lemma 3, we would have to use $\Pr[y \in I] \approx 1 - e^{-1/(4n^2)} \approx \frac{1}{4n^2}$ and thus would need the overall bound $0.001 + (4n^2 - 1) \cdot (1 - \varepsilon) - (1 - \frac{1}{4n^2}) < 1$. This means that ε should be $1 - \mathcal{O}(1/n^2)$, whereas the DISJ bound is only established for sufficiently small values of protocol error.

Acknowledgments. We thank Marc Fischlin for participating in the early stages of this work. We also thank the CRYPTO'18 (sub)reviewers for their valuable comments. Bauer was supported by the French ANR Project ANR-16-CE39-0002 EfTrEC. Farshim was supported by the European Research Council under the European Community's Seventh Framework Programme (FP7/2007-2013 Grant Agreement no. 339563 - CryptoCloud). Mazaheri was supported by the German Federal Ministry of Education and Research (BMBF) and by the Hessian State Ministry for Higher Education, Research and the Arts, within CRISP.

References

1. L. Babai, P. Frankl, and J. Simon. Complexity classes in communication complexity theory (preliminary version). In *27th FOCS*, pages 337–347, 1986. (Cited on pages 4, 5, and 14.)
2. Z. Bar-Yossef, T. S. Jayram, R. Kumar, and D. Sivakumar. An information statistics approach to data stream and communication complexity. In *43rd FOCS*, pages 209–218, 2002. (Cited on pages 4 and 14.)
3. B. Barak, M. Braverman, X. Chen, and A. Rao. How to compress interactive communication. In *42nd ACM STOC*, pages 67–76, 2010. (Cited on page 19.)
4. M. Bellare and P. Rogaway. Random oracles are practical: A paradigm for designing efficient protocols. In *ACM CCS 93*, pages 62–73, 1993. (Cited on page 1.)
5. D. J. Bernstein, T. Lange, and R. Niederhagen. Dual EC: A standardized back door. Cryptology ePrint Archive, Report 2015/767, 2015. http://eprint.iacr.org/2015/767. (Cited on page 2.)
6. D. Boneh and X. Boyen. On the impossibility of efficiently combining collision resistant hash functions. In *CRYPTO 2006*, pages 570–583, 2006. (Cited on page 2.)
7. R. Canetti, R. L. Rivest, M. Sudan, L. Trevisan, S. P. Vadhan, and H. Wee. Amplifying collision resistance: A complexity-theoretic treatment. In *CRYPTO 2007*, pages 264–283, 2007. (Cited on pages 5 and 25.)
8. A. Chattopadhyay and T. Pitassi. The story of set disjointness. *SIGACT News*, 41(3):59–85, 2010. (Cited on page 3.)
9. S. Checkoway, J. Maskiewicz, C. Garman, J. Fried, S. Cohney, M. Green, N. Heninger, R.-P. Weinmann, E. Rescorla, and H. Shacham. A systematic analysis of the juniper dual EC incident. In *ACM CCS 16*, pages 468–479, 2016. (Cited on page 2.)
10. S. Checkoway, R. Niederhagen, A. Everspaugh, M. Green, T. Lange, T. Ristenpart, D. J. Bernstein, J. Maskiewicz, H. Shacham, and M. Fredrikson. On the practical exploitability of dual EC in TLS implementations. In *23rd USENIX Security Symposium (USENIX Security 14)*, pages 319–335, 2014. (Cited on page 2.)
11. S. Coretti, Y. Dodis, S. Guo, and J. Steinberger. Random oracles and non-uniformity. Cryptology ePrint Archive, Report 2017/937, 2017. http://eprint.iacr.org/2017/937. (Cited on pages 7 and 9.)
12. I. Dinur. New attacks on the concatenation and XOR hash combiners. In *EUROCRYPT 2016, Part I*, pages 484–508, 2016. (Cited on page 2.)
13. Y. Dodis, S. Guo, and J. Katz. Fixing cracks in the concrete: Random oracles with auxiliary input, revisited. In *EUROCRYPT 2017, Part II*, pages 473–495, 2017. (Cited on page 7.)
14. A. Fiat and A. Shamir. How to prove yourself: Practical solutions to identification and signature problems. In *CRYPTO'86*, pages 186–194, 1987. (Cited on page 1.)
15. M. Fischlin and A. Lehmann. Security-amplifying combiners for collision-resistant hash functions. In *CRYPTO 2007*, pages 224–243, 2007. (Cited on page 2.)
16. M. Fischlin, A. Lehmann, and K. Pietrzak. Robust multi-property combiners for hash functions. *Journal of Cryptology*, 27(3):397–428, 2014. (Cited on page 2.)
17. O. Goldreich. *Foundations of Cryptography: Basic Tools*, volume 1. Cambridge University Press, Cambridge, UK, 2001. (Cited on pages 5 and 20.)
18. V. Guruswami and M. Cheraghchi. Set disjointness lower bound via product distribution. Scribes for Information theory and its applications in theory of computation, 2013. http://www.cs.cmu.edu/~venkatg/teaching/ITCS-spr2013/. (Cited on pages 4 and 15.)

19. J. J. Hoch and A. Shamir. On the strength of the concatenated hash combiner when all the hash functions are weak. In *ICALP 2008, Part II*, pages 616–630, 2008. (Cited on pages 2, 3, 6, and 13.)
20. A. Joux. Multicollisions in iterated hash functions. Application to cascaded constructions. In *CRYPTO 2004*, pages 306–316, 2004. (Cited on page 13.)
21. J. Katz, S. Lucks, and A. Thiruvengadam. Hash functions from defective ideal ciphers. In *CT-RSA 2015*, pages 273–290, 2015. (Cited on page 8.)
22. A. Kawachi, A. Numayama, K. Tanaka, and K. Xagawa. Security of encryption schemes in weakened random oracle models. In *PKC 2010*, pages 403–419, 2010. (Cited on pages 2 and 13.)
23. E. Kushilevitz and N. Nisan. *Communication complexity*. Cambridge University Press, 1997. (Cited on pages 3 and 13.)
24. A. Lehmann. *On the Security of Hash Function Combiners*. PhD thesis, TU Darmstadt, 2010. (Cited on page 12.)
25. G. Leurent and L. Wang. The sum can be weaker than each part. In *EUROCRYPT 2015, Part I*, pages 345–367, 2015. (Cited on pages 2 and 13.)
26. M. Liskov. Constructing an ideal hash function from weak ideal compression functions. In *SAC 2006*, pages 358–375, 2007. (Cited on pages 2 and 6.)
27. S. Lucks. A failure-friendly design principle for hash functions. In *ASIACRYPT 2005*, pages 474–494, 2005. (Cited on page 6.)
28. U. M. Maurer and S. Tessaro. A hardcore lemma for computational indistinguishability: Security amplification for arbitrarily weak PRGs with optimal stretch. In *TCC 2010*, pages 237–254, 2010. (Cited on pages 5 and 23.)
29. F. Mendel, C. Rechberger, and M. Schläffer. MD5 is weaker than weak: Attacks on concatenated combiners. In *ASIACRYPT 2009*, pages 144–161, 2009. (Cited on page 2.)
30. A. Mittelbach. Cryptophia's short combiner for collision-resistant hash functions. In *ACNS 13*, pages 136–153, 2013. (Cited on page 13.)
31. D. Moshkovitz and B. Barak. Communication complexity. Scribes for Advanced Complexity Theory, 2012. https://people.csail.mit.edu/dmoshkov/courses/adv-comp/. (Cited on pages 4 and 15.)
32. A. Numayama, T. Isshiki, and K. Tanaka. Security of digital signature schemes in weakened random oracle models. In *PKC 2008*, pages 268–287, 2008. (Cited on page 2.)
33. O. Reingold, L. Trevisan, and S. P. Vadhan. Notions of reducibility between cryptographic primitives. In *TCC 2004*, pages 1–20, 2004. (Cited on page 5.)
34. P. Rogaway and T. Shrimpton. Cryptographic hash-function basics: Definitions, implications, and separations for preimage resistance, second-preimage resistance, and collision resistance. In *FSE 2004*, pages 371–388, 2004. (Cited on page 10.)
35. D. R. Simon. Finding collisions on a one-way street: Can secure hash functions be based on general assumptions? In *EUROCRYPT'98*, pages 334–345, 1998. (Cited on pages 5, 19, and 23.)
36. D. Unruh. Random oracles and auxiliary input. In *CRYPTO 2007*, pages 205–223, 2007. (Cited on page 7.)
37. A. C.-C. Yao. Some complexity questions related to distributive computing(preliminary report). In *Proceedings of the Eleventh Annual ACM Symposium on Theory of Computing*, pages 209–213, 1979. (Cited on pages 3 and 13.)

A Image Uniformity

A.1 Image uniformity of combiners

We restate and prove Lemma 1 from Section 2 to bound the advantage of adversaries in the image uniformity game. Recall that we denote by $\mathcal{U}_S$ the uniform distribution over a set S. For (N, M) a pair of integers we let $\mathcal{U}_f^{\mathsf{p}}$ denote the distribution defined by $\mathcal{U}_f^{\mathsf{p}}(x) = |f^{-1}(x)|/N$, where $f : [N] \to [M]$ is a uniform function.

Lemma 1 (Combiner image uniformity). *Let $\mathsf{C}_t^{\mathsf{H}_1,\mathsf{H}_2} : \{0,1\}^n \to \{0,1\}^m$ be a combiner for $t \in \{|, \circ, \oplus\}$. Let $\mathsf{H} : \{0,1\}^n \to \{0,1\}^m$ be a hash function. Then*

$$\mathsf{Adv}^{\mathrm{iu}}_{\mathsf{C}_t^{\mathsf{H}_i}}(\mathcal{A}) \leq \mathbb{E}_{\mathsf{H}}\left[\Delta_{\mathsf{TV}}\big(\mathcal{U}_{\mathsf{Img}(\mathsf{H})}, \mathcal{U}_{\mathsf{H}}^{\mathsf{p}}\big)\right] + 2 \cdot p_t \ ,$$

where $p_| = p_\oplus = 0$ and $p_\circ \leq 2^{2n_1 - m_1}$ is the probability that $\mathsf{H}_1 : \{0,1\}^{n_1} \to \{0,1\}^{m_1}$ is not injective (i.e., it has at least one collision).

Proof. Let H_1, H_2, and H be random variables denoting uniform functions in $\mathrm{Fun}[n_1, m_1]$, $\mathrm{Fun}[n_2, m_2]$, and $\mathrm{Fun}[n, m]$ respectively. The distinguishing advantage of an adversary $\mathcal{A}$ in the image-uniformity game is, by definition, upper bounded as

$$\mathsf{Adv}^{\mathrm{iu}}_{\mathsf{C}_t^{\mathsf{H}_i}}(\mathcal{A}) \leq \Delta_{\mathsf{TV}}\big((\mathsf{H}_1, \mathsf{H}_2, \mathsf{C}_t^{\mathsf{H}_1,\mathsf{H}_2}(X)), (\mathsf{H}_1, \mathsf{H}_2, Y)\big) \ ,$$

where X is uniform over $\{0,1\}^n$ and Y is uniform over the image of $\mathsf{C}_t^{\mathsf{H}_1,\mathsf{H}_2}$.

For a uniform $\mathsf{H} : \{0,1\}^n \to \{0,1\}^m$ we claim that

$$\Delta_{\mathsf{TV}}\big((\mathsf{H}_1, \mathsf{H}_2, \mathsf{C}_t^{\mathsf{H}_1,\mathsf{H}_2}(X)), (\mathcal{S}_t(\mathsf{H}), \mathsf{H}(X))\big) \leq p_t$$
$$\Delta_{\mathsf{TV}}\big((\mathsf{H}_1, \mathsf{H}_2, Y), (\mathcal{S}_t(\mathsf{H}), Y)\big) \leq p_t \ ,$$

where the simulators $\mathcal{S}_t(\mathsf{H})$ output simulated H_1 and H_2 tables as follows.

Concatenation: The simulator outputs the left m_1 bits and the right m_2 bits of the outputs of H as the two hash functions respectively. It is easily seen that $(\mathsf{H}_1, \mathsf{H}_2)$ is identically distributed to $\mathcal{S}_|(\mathsf{H})$.

Xor: The simulator samples H_1 and defines $\mathsf{H}_2(x) := \mathsf{H}(x) \oplus \mathsf{H}_1(x)$ for all x. It is easily seen that $(\mathsf{H}_1, \mathsf{H}_2)$ is identically distributed to $\mathcal{S}_\oplus(\mathsf{H})$.

Cascade: The simulator samples H_1 and H_2. It then redefines H_2 to map all $\mathsf{H}_1(x)$ values to $\mathsf{H}(x)$. The latter results in a consistent random function unless $\mathsf{H}_1(x)$ is not injective. The probability of latter is, by definition, $p_\circ$.

It thus remains to bound

$$\Delta_{\mathsf{TV}}\big((\mathcal{S}_t(\mathsf{H}), \mathsf{H}(X)), (\mathcal{S}_t(\mathsf{H}), Y)\big) \ .$$

Since $\mathcal{S}_t(\mathsf{H})$ is a randomized transformation of H, we can simply bound $\Delta_{\mathsf{TV}}\big((\mathsf{H},\mathsf{H}(X)),(\mathsf{H},Y)\big)$. Note that $\mathsf{H}(X)$ is identical to $\mathcal{U}^{\mathsf{p}}_{\mathsf{H}}$ and Y is identical to $\mathcal{U}_{\mathsf{Img}(\mathsf{H})}$. We have

$$\begin{aligned}
\Delta_{\mathsf{TV}}\big((\mathsf{H},\mathcal{U}_{\mathsf{Img}(\mathsf{H})}),(\mathsf{H},\mathcal{U}^{\mathsf{p}}_{\mathsf{H}})\big) &= \frac{1}{2}\sum_{(h,y)}\big|\Pr[(\mathsf{H},\mathcal{U}_{\mathsf{Img}(\mathsf{H})})=(h,y)]-\Pr[(\mathsf{H},\mathcal{U}^{\mathsf{p}}_{\mathsf{H}})=(h,y)]\big| \\
&= \frac{1}{2}\sum_{(h,y)}\Pr[\mathsf{H}=h]\big|\Pr[h(x)=y]-\Pr[\mathcal{U}_{\mathsf{Img}(h)}=y]\big| \\
&= \frac{1}{2}\sum_{h}\Pr[\mathsf{H}=h]\sum_{y}\big|\Pr[h(x)=y]-\Pr[\mathcal{U}_{\mathsf{Img}(h)}=y]\big| \\
&= \frac{1}{2}\sum_{h}\Pr[\mathsf{H}=h]\Delta_{\mathsf{TV}}(\mathsf{H}(x),\mathcal{U}_{\mathsf{Img}(h)}) \\
&= \mathbb{E}_{h\in\mathrm{Fun}[n,m]}\left[\Delta_{\mathsf{TV}}(\mathcal{U}_{\mathsf{Img}(h)},\mathcal{U}^{\mathsf{p}}_{h})\right] \ .
\end{aligned}$$

For the upper bound on $p_\circ$, by Stirling's approximations and using $m_1 > 2n_1$ we have

$$\begin{aligned}
\Pr_{\mathsf{H}_1 \twoheadleftarrow \mathrm{Fun}[n_1,m_1]}[\mathsf{H}_1 \text{ is injective}] &= \frac{1}{(2^{m_1})^{2^{n_1}}}\cdot\frac{2^{m_1}!}{(2^{m_1}-2^{n_1})!} \\
&> e^{-2^{2n_1-m_1}} > 1-2^{2n_1-m_1} \ .
\end{aligned}$$

Thus

$$p_\circ := \Pr_{\mathsf{H}_1 \twoheadleftarrow \mathrm{Fun}[n_1,m_1]}[\mathsf{H}_1 \text{ not injective}] \leq 2^{2n_1-m_1} \ ,$$

which is negligible for $m_1 = (2+\epsilon)\cdot n_1$ with $\epsilon > 1$. We note that for $m_1 = 2n_1$ we asymptotically have that $p_\circ = 1 - 1/\sqrt{e}$.

□

A.2 Image uniformity of random functions

In the remaining of this section we bound the final expected statistical distance above. In the following we denote the set of all functions $f : [N] \to [M]$ by $[M]^N$, for integers $N, M \in \mathbb{N}$.

The expanding case. We start by treating expanding functions. The following lemma give us the expected image size for functions $f \in [M]^N$.

Lemma 8. *Let $N, M \in \mathbb{N}$. Then the expected image size (the expected number of non-empty bins) is*

$$\mathbb{E}_{f \twoheadleftarrow [M]^N}[|\mathsf{Img}(f)|] = M - M\cdot(1-\frac{1}{M})^N \ .$$

Proof. Below, all probabilities and expectations are taken over a random choice of $f \twoheadleftarrow [M]^N$.

$$\begin{aligned}
\mathbb{E}[|\mathsf{Img}(f)|] &= \mathbb{E}[|\{y \in [M] : \exists x \in [N] : f(x) = y\}|] \\
&= \sum_{y \in [M]} \Pr[\exists x \in [N] : f(x) = y] \\
&= M \cdot (1 - \Pr[\forall x \in [N] : f(x) \neq y]) \\
&= M \cdot (1 - (1 - 1/M)^N) \\
&= M - M \cdot (1 - 1/M)^N
\end{aligned}$$

In the second equality above we have used the linearity of expectation. □

We now bound the expected statistical distance for expanding functions in the following lemma.

Lemma 9. *Let N, M some integers, let $1 \leq \alpha_N < N$. Then*

$$\mathbb{E}_{f \twoheadleftarrow [M]^N} \left[\Delta_{\mathsf{TV}}(\mathcal{U}_{\mathsf{Img}(f)}, \mathcal{U}_f^{\mathsf{p}}) \right] \leq \Pr[N - |f([N])| \geq \alpha_N] + \frac{\alpha_N}{N} \ .$$

Proof. Let $E_{\alpha_N} := \{f \in [M]^N : N - |f([N])| < \alpha_N\}$. We have

$$\mathbb{E}_{f \twoheadleftarrow [M]^N} \left[\Delta_{\mathsf{TV}}(\mathcal{U}_{\mathsf{Img}(f)}, \mathcal{U}_f^{\mathsf{p}}) \right] \leq \Pr[N - |f([N])| \geq \alpha_N] + \mathbb{E}_{f \twoheadleftarrow E_{\alpha_N}} [\Delta_{\mathsf{TV}}(\mathcal{U}_{\mathsf{Img}(f)}, \mathcal{U}_f^{\mathsf{p}})] \ .$$

In order to upper bound $\Delta_{\mathsf{TV}}(\mathcal{U}_{\mathsf{Img}(f)}, \mathcal{U}_f^{\mathsf{p}})$ for each $f \in E_{\alpha_N}$ we define

$$S^- := \{i \in [M] : \mathcal{U}_{\mathsf{Img}(f)}(i) - \mathcal{U}_f^{\mathsf{p}}(i) < 0\} \ .$$

Then, by the symmetry of statistical distance, and the fact that $|f([N])| \leq N$ we have

$$\begin{aligned}
\Delta_{\mathsf{TV}}(\mathcal{U}_{\mathsf{Img}(f)}, \mathcal{U}_f^{\mathsf{p}}) &\leq \sum_{i \in S^-} \left(\frac{|f^{-1}(i)|}{N} - \frac{1}{N} \right) \\
&\leq \sum_{i \in f([N])} \left(\frac{|f^{-1}(i)|}{N} - \frac{1}{N} \right) \\
&\leq \frac{N - |f([N])|}{N} \leq \frac{\alpha_N}{N} \ .
\end{aligned}$$

We have also used the facts that the summands are positive. Note also that $\sum_{i \in f([N])} |f^{-1}(i)| = N$ and $\sum_{i \in f([N])} 1 = |f([N])|$. This concludes the proof of the claim. □

For large M,

$$\ln\left(1 - \frac{1}{M}\right) = -\frac{1}{M} - \frac{1}{2M^2} + \mathcal{O}\left(\frac{1}{M^3}\right) \ .$$

Hence

$$\left(1 - \frac{1}{M}\right)^N = \exp\left(N \cdot \ln\left(1 - \frac{1}{M}\right)\right) = \exp\left(-\frac{N}{M} - \frac{N}{2M^2} + \mathcal{O}\left(\frac{N}{M^3}\right)\right) \ .$$

Corollary 3. *Let $0 < \gamma < 1$ and C be a positive constant. Suppose $N = C \cdot M^{\gamma}$. Then $\mathbb{E}_{f \twoheadleftarrow [M]^N}(\Delta_{\mathsf{TV}}(\mathcal{U}_{\mathsf{Img}(f)}, \mathcal{U}_f^{\mathsf{p}}))$ is negligible.*

Proof. We set $\alpha_N := M^{(1-\gamma)/2}\left(N - M + M(1 - \frac{1}{M})^N\right)$. By construction, and using the expected number of collisions and Markov's inequality,[11] we have

$$\Pr[N - |f([N])| \geq \alpha_N] \leq 1/M^{(1-\gamma)/2} \ .$$

Thus the first term in the bound shown in the claim above is negligible. It remains to show that α_N/N is also negligible. We have

$$\begin{aligned}
\alpha_N &= M^{(1-\gamma)/2}\left(CM^{\gamma} - M + M\exp\left(-C \cdot M^{\gamma-1} - C/2 \cdot M^{\gamma-2} + \mathcal{O}\left(M^{\gamma-3}\right)\right)\right) \\
&= M^{(1-\gamma)/2}\left(C \cdot M^{\gamma} - M + M\left(1 - C \cdot M^{\gamma-1} + \mathcal{O}\left(M^{2\gamma-2}\right)\right)\right) \\
&= M^{(1-\gamma)/2}\mathcal{O}\left(M^{2\gamma-1}\right) = \mathcal{O}\left(M^{3/2\cdot\gamma-1/2}\right) \ .
\end{aligned}$$

Thus $\alpha_N/N = \mathcal{O}(M^{(\gamma-1)/2})$, which is negligible for $\gamma < 1$. □

The length-preserving case. We now deal with the case where the function is either length preserving or only slightly expanding.

Lemma 10. *Suppose $N = CM$ with $C \leq 1$, and $0 < \delta < 1$. Set $\mu_1 := M(1 - \frac{1}{e^C})$ and $\mu_2 := \frac{CM}{e^C}$. Then*

$$\begin{aligned}
\mathbb{E}_{f \twoheadleftarrow [M]^N}[\Delta_{\mathsf{TV}}(\mathcal{U}_{\mathsf{Img}(f)}, \mathcal{U}_f^{\mathsf{p}})] \leq \Pr\left[|f([N])| < (1-\delta)\,\mu_1\right] + \\
+ \Pr\left[\left|\{i \in [M] : \left|f^{-1}(i)\right| = 1\}\right| > (1+\delta)\,\mu_2\right] \\
+ \frac{(1+\delta)\left(C - (1-\delta)\left(1 - \frac{1}{e^C}\right)\right)}{e^C\,(1-\delta)\left(1 - \frac{1}{e^C}\right)} \ ,
\end{aligned}$$

where for negligible δ, the first two terms are negligible and the last term is $\frac{1}{e^C} \cdot \left(\frac{C}{(1 - \frac{1}{e^C})} - 1\right)$.

Proof. Let $F := F_1 \cap F_2$, where

$$\begin{aligned}
F_1 &:= \{f \in [M]^N : |f([N])| \geq (1-\delta)\,\mu_1\} \\
F_2 &:= \{f \in [M]^N : \left|\{i \in [M] : \left|f^{-1}(i)\right| = 1\}\right| \leq (1+\delta)\,\mu_2\} \ .
\end{aligned}$$

Then

$$\begin{aligned}
\mathbb{E}_{f \twoheadleftarrow [M]^N}[\Delta_{\mathsf{TV}}(\mathcal{U}_{\mathsf{Img}(f)}, \mathcal{U}_f^{\mathsf{p}})] &\leq \Pr_{f \twoheadleftarrow [M]^N}[f \notin F] + \mathbb{E}_{f \twoheadleftarrow F}\left[\Delta_{\mathsf{TV}}\left(\mathcal{U}_{\mathsf{Img}(f)}, \mathcal{U}_f^{\mathsf{p}}\right)\right] \\
&\leq \Pr\left[|f([N])| < (1-\delta)\,\mu_1\right] + \Pr\left[\left|\{i \in [M] : \left|f^{-1}(i)\right| = 1\}\right| > (1+\delta)\,\mu_2\right] \\
&\quad + \mathbb{E}_{f \twoheadleftarrow F}\left[\Delta_{\mathsf{TV}}(\mathcal{U}_{\mathsf{Img}(f)}, \mathcal{U}_f^{\mathsf{p}})\right] \ .
\end{aligned}$$

[11] Recall this states that $\Pr[X \geq t\,\mathbb{E}[X]] \leq 1/t$ for positive X.

For each $f \in F$ define

$$S_f^- := \{i \in [M] : \mathcal{U}_{\mathsf{Img}(f)}(i) - \mathcal{U}_f^{\mathsf{p}}(i) < 0\} \ .$$

For $f \in F$ we have $|f([N])| \geq (1-\delta)\,\mu_1$ and for $i \in S_f^-$ we have $\mathcal{U}_{\mathsf{Img}(f)}(i) < \mathcal{U}_f^{\mathsf{p}}(i)$. Thus,

$$\frac{\left|f^{-1}(i)\right|}{N} < \frac{1}{|f([N])|} \leq \frac{1}{(1-\delta)\,\mu_1} \ ,$$
$$\left|f^{-1}(i)\right| \leq \frac{C}{\left(1-\frac{1}{e^C}\right)} \leq \frac{1}{\left(1-\frac{1}{e}\right)} < 2 \ .$$

The penultimate inequality is due to the fact that the map $C \mapsto \frac{C}{\left(1-\frac{1}{e^C}\right)}$ has a positive derivative and that we have assumed $C \leq 1$. Hence $\left|f^{-1}(i)\right| < 2$ and because $i \in f([N])$, we can deduce $\left|f^{-1}(i)\right| = 1$ (since $\left|f^{-1}(i)\right|$ is an integer). Hence $S_f^- \subseteq \{i \in [M] : \left|f^{-1}(i)\right| = 1\}$, which allows us to upper bound the cardinality of S_f^- by $(1+\delta)\,\mu_2$ (since $f \in F \subseteq F_2$). Now

$$\begin{aligned}
\Delta_{\mathsf{TV}}\left(\mathcal{U}_{\mathsf{Img}(f)}, \mathcal{U}_f^{\mathsf{p}}\right) &= \sum_{i \in S_f^-} \left(\frac{1}{|f(([N])|} - \frac{\left|f^{-1}(i)\right|}{N}\right) \\
&\leq \sum_{i \in S_f^-} \left(\frac{1}{(1-\delta)\,M\left(1-\frac{1}{e^C}\right)} - \frac{1}{N}\right) \\
&\leq \frac{(1+\delta)\,CM}{e^C}\left(\frac{1}{(1-\delta)\,M\left(1-\frac{1}{e^C}\right)} - \frac{1}{CM}\right) \\
&= \frac{(1+\delta)\left(C - (1-\delta)\left(1-\frac{1}{e^C}\right)\right)}{e^C\,(1-\delta)\left(1-\frac{1}{e^C}\right)} \ .
\end{aligned}$$

If we set $\delta := N^{-1/4}$ the (expectation of the) above term is asymptotically upper bounded by $\frac{1}{e^C} \cdot \left(\frac{C}{\left(1-\frac{1}{e^C}\right)} - 1\right)$ as N grows.

We use can also the Chernoff bounds to upper bound the two remaining probabilities. We note that

$$\mathbb{E}_{f \twoheadleftarrow [M]^N}\left[|f([N])|\right] = M - M\left(1-\frac{1}{M}\right)^{CM} \geq M(1-\frac{1}{e^C}) = \mu_1 \ ,$$

and that

$$\begin{aligned}
\mathbb{E}_{f \twoheadleftarrow [M]^N}\left[\left|\{i \in [M] : \left|f^{-1}(i)\right| = 1\}\right|\right] &= \sum_{i \in [M]} \Pr_{f \twoheadleftarrow [M]^N}\left(\left|f^{-1}(i)\right| = 1\right) \\
&= M \cdot \binom{CM}{1}\left(1-\frac{1}{M}\right)^{CM-1}\frac{1}{M} \\
&\leq \frac{CM}{e^C} = \mu_2 \ .
\end{aligned}$$

Thus μ_1 lower bounds the first exception and μ_2 upper bounds the second expectation. Using these facts and Chernoff we obtain

$$\Pr_{f \twoheadleftarrow [M]^N}\left[\left|f\left([N]\right)\right| < (1-\delta)\,\mu_1\right] \leq e^{-\frac{1}{2}\delta^2 \cdot M\left(1-\left(1-\frac{1}{M}\right)^{CM}\right)} \leq e^{-\frac{1}{2C}\cdot\sqrt{N}\left(1-\frac{1}{e^C}\right)}$$

$$\Pr_{f \twoheadleftarrow [M]^N}\left[\left|\left\{i \in [M] : \left|f^{-1}\left(i\right)\right| = 1\right\}\right| > (1+\delta)\,\mu_2\right] \leq e^{-\frac{1}{3}\delta^2 \cdot CM\left(1-\frac{1}{M}\right)^{CM-1}} \leq e^{-\frac{1}{3eC}\cdot\sqrt{N}}$$

where we used the facts that $\left(1-\frac{1}{M}\right)^{CM} \leq \frac{1}{e^C}$ and $\left(1-\frac{1}{M}\right)^{CM-1} \geq \frac{1}{e^C}$. □

The compressing case. We now deal with the case of compressing functions $f : [N] \to [M]$.

Lemma 11. *Let N, M be integers and $\delta > 0$. Then*

$$\mathbb{E}_{f \twoheadleftarrow [M]^N}\left[\Delta_{\mathsf{TV}}(\mathcal{U}_{\mathsf{Img}(f)}, \mathcal{U}_f^{\mathsf{p}})\right] \leq \Pr_{f \twoheadleftarrow [M]^N}\left[\exists i \in [M] : \left|\left|f^{-1}(i)\right| - \frac{N}{M}\right| \geq \delta\frac{N}{M}\right] + \Pr_{f \twoheadleftarrow [M]^N}\left[f\left([N]\right) \neq [M]\right] + \frac{\delta}{2}\ .$$

Proof. Let

$$F := \{f \in [M]^N : f\left([N]\right) = [M] \wedge \forall i \in [M]\left|\left|f^{-1}(i)\right| - \frac{N}{M}\right| < \delta\frac{N}{M}\}\ .$$

Then

$$\mathbb{E}_{f \twoheadleftarrow [M]^N}\left[\Delta_{\mathsf{TV}}(\mathcal{U}_{\mathsf{Img}(f)}, \mathcal{U}_f^{\mathsf{p}})\right] \leq \Pr_{f \twoheadleftarrow [M]^N}\left[f \notin F\right] + \mathbb{E}_{f \twoheadleftarrow F}\left[\Delta_{\mathsf{TV}}(\mathcal{U}_{\mathsf{Img}(f)}, \mathcal{U}_f^{\mathsf{p}})\right]\ .$$

Let $f \in F$. Then

$$\begin{aligned}\Delta_{\mathsf{TV}}(\mathcal{U}_{\mathsf{Img}(f)}, \mathcal{U}_f^{\mathsf{p}}) &\leq \frac{1}{2}\sum_{i\in[M]}\left|\frac{\left|f^{-1}(i)\right|}{N} - \frac{1}{M}\right| \\ &\leq \frac{1}{2N}\sum_{i\in[M]}\left|\left|f^{-1}(i)\right| - \frac{N}{M}\right| \\ &\leq \frac{1}{2N}\sum_{i\in[M]}\delta\frac{N}{M} \leq \frac{\delta}{2}\ .\end{aligned}$$

□

Corollary 4. *Suppose $N = M^\gamma$ with $\gamma > 1$. Then $\mathbb{E}_{f \twoheadleftarrow [M]^N}\left[\Delta_{\mathsf{TV}}(\mathcal{U}_{\mathsf{Img}(f)}, \mathcal{U}_f^{\mathsf{p}})\right]$ is negligible.*

Proof. Let $\delta := M^{(1-\gamma)/4}$. Note that δ is negligible. Furthermore,

$$\begin{aligned}\Pr\left[f\left([N]\right) \neq [M]\right] &\leq \Pr\left[\exists i \in [M] : f^{-1}(i) = \emptyset\right] \\ &\leq M \Pr\left[f^{-1}(0) = \emptyset\right] \\ &\leq M\left(1 - \frac{1}{M}\right)^N \leq M e^{-M^{\gamma-1}} \ ,\end{aligned}$$

which is negligible. We now upper bound the first term in the claim above.

$$\begin{aligned}&\Pr_{f \twoheadleftarrow [M]^N}\left[\exists i \in [M] : \left|\left|f^{-1}(i)\right| - \frac{N}{M}\right| \geq \delta\frac{N}{M}\right] \\ &\qquad \leq \sum_{i=0}^{M-1} \Pr_{f \twoheadleftarrow [M]^N}\left[\left|\left|f^{-1}(i)\right| - \frac{N}{M}\right| \geq \delta\frac{N}{M}\right] \\ &\qquad \leq M \Pr_{f \twoheadleftarrow [M]^N}\left[\left|\left|f^{-1}(0)\right| - \frac{N}{M}\right| \geq \delta\frac{N}{M}\right] \ .\end{aligned}$$

By the Chernoff bounds we have

$$\Pr_{f \twoheadleftarrow [M]^N}\left[\exists i \in [M] : \left|\left|f^{-1}(i)\right| - \frac{N}{M}\right| \geq \delta\frac{N}{M}\right] \leq 2e^{-\frac{\delta^2 N}{3M}} = 2e^{-\frac{M^{(\gamma-1)/2}}{3}} \ ,$$

which when multiplied by M remains negligible. □

B Proofs of rPre + IU $\implies$ OW and oPRG + IU $\implies$ PRG

In this section we restate and prove Lemmas 2 and 3 from Section 2.

Lemma 2 (rPre + IU $\implies$ OW)**.** *Let* $\mathsf{C}^{\mathsf{H}_i} \in \mathrm{Fun}[n, m]$ *be a construction in the* 2*-BRO model. Then for any adversary* $\mathcal{A}$ *against the one-way security of* $\mathsf{C}^{\mathsf{H}_i}$*, there is an adversary* $\mathcal{B}$ *against the image uniformity and an adversary* $\mathcal{C}$ *against the* rPre *security of* $\mathsf{C}^{\mathsf{H}_i}$*, all in the* 2*-BRO model and using identical backdoor functionalities, such that*

$$\mathsf{Adv}^{\mathrm{ow}}_{\mathsf{C}^{\mathsf{H}_i}}(\mathcal{A}) \leq \mathsf{Adv}^{\mathrm{iu}}_{\mathsf{C}^{\mathsf{H}_i}}(\mathcal{B}) + \frac{1}{\alpha} \cdot \mathsf{Adv}^{\mathrm{rpre}}_{\mathsf{C}^{\mathsf{H}_i}}(\mathcal{C}) - \frac{1-\alpha}{\alpha} \ ,$$

where $\alpha := \Pr[y \in \mathsf{Img}(\mathsf{C}^{\mathsf{H}_i})]$ *over a random choice of* $y \in \{0,1\}^m$ *and* H_i*.*

Proof. Let OW-RI be a game which differs from the standard OW game in that in OW-RI the challenge y is chosen randomly from the set of possible images $\mathsf{Img}(\mathsf{C}^{\mathsf{H}_i})$ rather than indirectly as the image of a randomly chosen $x \in \{0,1\}^n$. Any difference in the advantage of an adversary $\mathcal{A}$ in these two games is bounded by the advantage of an adversary $\mathcal{B}$ in the IU game, i.e.,

$$\mathsf{Adv}^{\mathrm{ow}}_{\mathsf{C}^{\mathsf{H}_i}}(\mathcal{A}) - \mathsf{Adv}^{\text{ow-ri}}_{\mathsf{C}^{\mathsf{H}_i}}(\mathcal{A}) \leq \mathsf{Adv}^{\mathrm{iu}}_{\mathsf{C}^{\mathsf{H}_i}}(\mathcal{B}) \ .$$

We can now build an adversary $\mathcal{C}$ against rPre security of $\mathsf{C}^{\mathsf{H}_i}$ as follows. On input $y \in \{0,1\}^m$, algorithm $\mathcal{C}^{\mathrm{BD}_i}$ simply runs $\mathcal{A}^{\mathrm{BD}_i}(y)$, while using its own backdoor oracles to answer $\mathcal{A}$'s queries. After $\mathcal{A}$ terminates with some x, adversary $\mathcal{C}$ outputs x if $\mathsf{C}^{\mathsf{H}_i}(x) = y$ and $\perp$ otherwise.

Let a and c be the outputs of $\mathcal{A}^{\mathrm{BD}_i}(y)$ and $\mathcal{C}^{\mathrm{BD}_i}(y)$ respectively. Let $I := \mathsf{Img}(\mathsf{C}^{\mathsf{H}_i})$, i.e., the image of $\mathsf{C}^{\mathsf{H}_i}$. Below we analyze $\mathcal{C}$'s advantage in the game rPre where all probabilities are over $y \twoheadleftarrow \{0,1\}^m$ and random choices of H_i.

$$\begin{aligned}
\mathsf{Adv}^{\mathrm{rpre}}_{\mathsf{C}^{\mathsf{H}_i}}(\mathcal{C}) &= \Pr[\mathsf{C}^{\mathsf{H}_i}(c) = y \wedge y \in I] + \Pr[c = \perp \wedge y \notin I] \\
&= \Pr[\mathsf{C}^{\mathsf{H}_i}(a) = y \wedge y \in I] + \Pr[\mathsf{C}^{\mathsf{H}_i}(a) \neq y \wedge y \notin I] \\
&= \Pr[\mathsf{C}^{\mathsf{H}_i}(a) = y \mid y \in I] \cdot \Pr[y \in I] + \Pr[\mathsf{C}^{\mathsf{H}_i}(a) \neq y \wedge y \notin I] \\
&= \mathsf{Adv}^{\text{ow-ri}}_{\mathsf{C}^{\mathsf{H}_i}}(\mathcal{A}) \cdot \Pr[y \in I] + \Pr[\mathsf{C}^{\mathsf{H}_i}(a) \neq y \mid y \notin I] \cdot \Pr[y \notin I] \\
&= \mathsf{Adv}^{\text{ow-ri}}_{\mathsf{C}^{\mathsf{H}_i}}(\mathcal{A}) \cdot \alpha + 1 \cdot (1 - \alpha) \\
&\geq \mathsf{Adv}^{\mathrm{ow}}_{\mathsf{C}^{\mathsf{H}_i}}(\mathcal{A}) \cdot \alpha - \mathsf{Adv}^{\mathrm{iu}}_{\mathsf{C}^{\mathsf{H}_i}}(\mathcal{B}) \cdot \alpha + (1 - \alpha)
\end{aligned}$$

We get

$$\mathsf{Adv}^{\mathrm{ow}}_{\mathsf{C}^{\mathsf{H}_i}}(\mathcal{A}) \leq \mathsf{Adv}^{\mathrm{iu}}_{\mathsf{C}^{\mathsf{H}_i}}(\mathcal{B}) + \frac{1}{\alpha} \cdot \mathsf{Adv}^{\mathrm{rpre}}_{\mathsf{C}^{\mathsf{H}_i}}(\mathcal{C}) - \frac{1-\alpha}{\alpha} \ .$$

□

As long as $\mathsf{Adv}^{\mathrm{rpre}}_{\mathsf{C}^{\mathsf{H}_i}}(\mathcal{C}) < 1 - \alpha \cdot \mathsf{Adv}^{\mathrm{iu}}_{\mathsf{C}^{\mathsf{H}_i}}(\mathcal{B})$ (and is not too close to 1), we get $\mathsf{Adv}^{\mathrm{ow}}_{\mathsf{C}^{\mathsf{H}_i}}(\mathcal{A}) < 1$, which means weak rPre with advantage smaller than $1 - \alpha \cdot \mathsf{Adv}^{\mathrm{iu}}_{\mathsf{C}^{\mathsf{H}_i}}(\mathcal{B})$ translates to weak one-way security. Note that a non-negligible bound on IU advantage would be sufficient for weak OW security as long as rPre advantage is sufficiently small.

Lemma 3 (oPRG + IU $\implies$ PRG)**.** *Let $\mathsf{C}^{\mathsf{H}_i} \in \mathrm{Fun}[n,m]$ be a construction in the 2-BRO model which is expanding with $m - n \geq 0.53$. Then for any adversary $\mathcal{A}$ against the PRG security of $\mathsf{C}^{\mathsf{H}_i}$, there is an adversary $\mathcal{B}$ against the image uniformity and an adversary $\mathcal{C}$ against the* oPRG *security of $\mathsf{C}^{\mathsf{H}_i}$, both in the 2-BRO model and using identical backdoor functionalities, such that*

$$\mathsf{Adv}^{\mathrm{prg}}_{\mathsf{C}^{\mathsf{H}_i}}(\mathcal{A}) \leq \mathsf{Adv}^{\mathrm{iu}}_{\mathsf{C}^{\mathsf{H}_i}}(\mathcal{B}) + \frac{1-\alpha}{\alpha} \cdot \mathsf{Adv}^{\mathrm{oprg}}_{\mathsf{C}^{\mathsf{H}_i}}(\mathcal{C}) - (1 - \alpha) \ ,$$

where $\alpha := \Pr[y \in \mathsf{Img}(\mathsf{C}^{\mathsf{H}_i})]$ over a random choice of $y \in \{0,1\}^m$ and H_i.

Proof. Let PRG-RI be a game which differs from the standard PRG game in that in PRG-RI if the secret bit is 1 the challenge y is chosen randomly from the set of possible images $\mathsf{Img}(\mathsf{C}^{\mathsf{H}_i})$ and not indirectly as the image of a randomly chosen $x \in \{0,1\}^n$. Any difference in the advantage of an adversary $\mathcal{A}$ in these two games is bounded by the advantage of an adversary $\mathcal{B}$ in the IU game, i.e.,

$$\mathsf{Adv}^{\mathrm{prg}}_{\mathsf{C}^{\mathsf{H}_i}}(\mathcal{A}) - \mathsf{Adv}^{\text{prg-ri}}_{\mathsf{C}^{\mathsf{H}_i}}(\mathcal{A}) \leq \mathsf{Adv}^{\mathrm{iu}}_{\mathsf{C}^{\mathsf{H}_i}}(\mathcal{B}) \ .$$

Note that we use image uniformity only once above, since in case $b = 1$ the PRG and PRG-RI games are identcial.

Next we build an adversary $\mathcal{C}$ against the oPRG security of $\mathsf{C}^{\mathsf{H}_i}$ as follows. On input $y \in \{0,1\}^m$, algorithm $\mathcal{C}$ simply runs $\mathcal{A}^{\mathrm{BD}_i}(y)$, while using its own backdoor oracles to answer $\mathcal{A}$'s queries. After $\mathcal{A}$ terminates with a guess bit, $\mathcal{C}$ outputs the same bit as its own guess.

Let a and c be shorthand for $\mathcal{A}^{\mathrm{BD}_i}(y) = 1$ and $\mathcal{C}^{\mathrm{BD}_i}(y) = 1$, respectively. Let $I := \mathsf{Img}(\mathsf{C}^{\mathsf{H}_i})$. Below we analyze $\mathcal{C}$'s advantage in the game oPRG, where all probabilities are over $y \twoheadleftarrow \{0,1\}^m$ and random choices of functions H_i.

$$\begin{aligned}
\mathsf{Adv}^{\mathrm{oprg}}_{\mathsf{C}^{\mathsf{H}_i}}(\mathcal{C}) &= \Pr[c \wedge y \in I] + \Pr[\neg c \wedge y \notin I] \\
&= \Pr[a \wedge y \in I] + \Pr[\neg a \wedge y \notin I] \\
&= \Pr[a \mid y \in I] \cdot \Pr[y \in I] + \Pr[\neg a \mid y \notin I] \cdot \Pr[y \notin I] \\
&= \Pr[a \mid y \in I] \cdot \alpha + \Pr[\neg a \mid y \notin I] \cdot (1 - \alpha)
\end{aligned}$$

Let $b_{\text{prg-ri}}$ be the challenge bit chosen in the PRG-RI game. We can now write the advantage of $\mathcal{A}$ in the intermediate PRG-RI game as below.

$$\begin{aligned}
\mathsf{Adv}^{\text{prg-ri}}_{\mathsf{C}^{\mathsf{H}_i}}(\mathcal{A}) &= 2 \cdot \Pr[a = b_{\text{prg-ri}}] - 1 \\
&= \Pr[a \mid y \in I] + \Pr[\neg a] - 1 \\
&= \Pr[a \mid y \in I] + \Pr[\neg a \mid y \in I] \cdot \Pr[y \in I] + \Pr[\neg a \mid y \notin I] \cdot \Pr[y \notin I] - 1 \\
&= \Pr[a \mid y \in I] + \Pr[\neg a \mid y \in I] \cdot \alpha + \Pr[\neg a \mid y \notin I] \cdot (1 - \alpha) - 1 \\
&= \Pr[a \mid y \in I] + (1 - \Pr[a \mid y \in I]) \cdot \alpha + \Pr[\neg a \mid y \notin I] \cdot (1 - \alpha) - 1 \\
&= \Pr[a \mid y \in I] \cdot (1 - \alpha) + \Pr[\neg a \mid y \notin I] \cdot (1 - \alpha) - (1 - \alpha)
\end{aligned}$$

Putting the above together and assuming that $1 - \alpha \geq 1/2$ (which follows from $s := m - n \geq 0.53$ and $\alpha = 1 - (1 - 1/2^m)^{2^n} \approx 1 - e^{-2^{-s}}$) we obtain

$$\begin{aligned}
&\alpha \cdot \mathsf{Adv}^{\text{prg-ri}}_{\mathsf{C}^{\mathsf{H}_i}}(\mathcal{A}) - (1 - \alpha) \cdot \mathsf{Adv}^{\mathrm{oprg}}_{\mathsf{C}^{\mathsf{H}_i}}(\mathcal{C}) \\
&\qquad = ((1 - \alpha) - 2(1 - \alpha)^2) \cdot \Pr[\neg a \mid y \notin I] + (1 - \alpha)^2 - (1 - \alpha) \\
&\qquad \leq (1 - \alpha)^2 - (1 - \alpha) \ .
\end{aligned}$$

The equality follows after some algebra and the inequality uses the fact that when $x^2 - 2x \leq 0$ for $x \geq 1/2$. Rearranging we get

$$\mathsf{Adv}^{\text{prg-ri}}_{\mathsf{C}^{\mathsf{H}_i}}(\mathcal{A}) \leq \frac{1 - \alpha}{\alpha} \cdot \mathsf{Adv}^{\mathrm{oprg}}_{\mathsf{C}^{\mathsf{H}_i}}(\mathcal{C}) - (1 - \alpha) \ ,$$

and overall we conclude that

$$\mathsf{Adv}^{\mathrm{prg}}_{\mathsf{C}^{\mathsf{H}_i}}(\mathcal{A}) \leq \mathsf{Adv}^{\mathrm{iu}}_{\mathsf{C}^{\mathsf{H}_i}}(\mathcal{B}) + \frac{1 - \alpha}{\alpha} \cdot \mathsf{Adv}^{\mathrm{oprg}}_{\mathsf{C}^{\mathsf{H}_i}}(\mathcal{C}) - (1 - \alpha) \ .$$

□

When $\mathsf{Adv}^{\mathrm{oprg}}_{\mathsf{C}^{\mathsf{H}_i}}(\mathcal{C}) < (2 - \alpha - \mathsf{Adv}^{\mathrm{iu}}_{\mathsf{C}^{\mathsf{H}_i}}(\mathcal{B})) \cdot \alpha/(1 - \alpha)$ (and is not too close to 1), we get an overall bound that is less than 1 and hence weak PRG security.

C Set-Disjointness

C.1 Information-theoretic inequalities

For two random variables X and Y over a common support D we define their statistical distance (or total variance) as usual via

$$\Delta_{\mathsf{TV}}(X,Y) := \frac{1}{2}\sum_{z\in D} |\Pr[X=z] - \Pr[Y=z]| \ .$$

The Hellinger distance between random variables X and Y is

$$\Delta_{\mathsf{Hel}}(X,Y) := \sqrt{1 - \sum_{z\in D} \sqrt{\Pr[X=z]\Pr[Y=z]}} \ .$$

The statistical and Hellinger distance are related via

$$\Delta^2_{\mathsf{Hel}}(X,Y) \le \Delta_{\mathsf{TV}}(X,Y) \le \sqrt{2}\cdot\Delta_{\mathsf{Hel}}(X,Y) \ .$$

To see this, let $p_z := \Pr[X=z]$ and $q_z := \Pr[Y=z]$. For the first inequality, we proceed as follows, where all the sums are over $z \in D$:

$$\begin{aligned}
\Delta^2_{\mathsf{Hel}}(X,Y) &= 1 - \sum\sqrt{p_z q_z} = 1/2\cdot\sum(p_z+q_z) - \sum\sqrt{p_z q_z} \\
&= 1/2\cdot\sum(\sqrt{p_z}-\sqrt{q_z})^2 \\
&= 1/2\cdot\sum|\sqrt{p_z}-\sqrt{q_z}||\sqrt{p_z}-\sqrt{q_z}| \\
&\le 1/2\cdot\sum|\sqrt{p_z}-\sqrt{q_z}||\sqrt{p_z}+\sqrt{q_z}| \\
&\le 1/2\cdot\sum|p_z - q_z| \\
&= \Delta_{\mathsf{TV}}(X,Y) \ .
\end{aligned}$$

For the second inequality, again taking all the sums over $z \in D$, we have

$$\begin{aligned}
\Delta^2_{\mathsf{TV}}(X,Y) &= 1/4\cdot\Big(\sum|p_z-q_z|\Big)^2 \\
&= 1/4\cdot\Big(\sum|\sqrt{p_z}-\sqrt{q_z}||\sqrt{p_z}+\sqrt{q_z}|\Big) \\
&\le 1/4\cdot\Big(\sum(\sqrt{p_z}-\sqrt{q_z})^2\Big)\Big(\sum(\sqrt{p_z}+\sqrt{q_z})^2\Big) \\
&= 1/2\cdot\Delta^2_{\mathsf{Hel}}(X,Y)\cdot\Big(2+2\sum\sqrt{p_z q_z}\Big) \\
&= \Delta^2_{\mathsf{Hel}}(X,Y)\cdot(2-\Delta^2_{\mathsf{Hel}}(X,Y)) \\
&\le 2\cdot\Delta^2_{\mathsf{Hel}}(X,Y) \ .
\end{aligned}$$

The first inequality (3rd line above) is an application of the Cauchy–Schwarz inequality, which that states $\left(\sum uv\right)^2 \le \sum u^2 \sum v^2$.

The Kullback–Leibler (KL) divergence between probability distributions $P(z)$ and $Q(z)$ is

$$D_{\mathrm{KL}}(P(z) \,\|\, Q(z)) := -\sum_{z \in D} P(z) \cdot \log \frac{Q(z)}{P(z)} \ .$$

Lemma 14. *Let Π be a random variable and $x_1, \ldots, x_N, y_1, \ldots, y_N$ be independent random variables. Then*

$$I(x_1, \ldots, x_N, y_1, \ldots, y_N; \Pi) \geq \sum_{i=1}^{N} I(x_i, y_i; \Pi) \ .$$

Proof.

$$\begin{aligned}
&I(x_1, \ldots, x_N, y_1, \ldots, y_N; \Pi) \\
&\quad = H(x_1, \ldots, x_N, y_1, \ldots, y_N) + H(\Pi) - H(x_1, \ldots, x_N, y_1, \ldots, y_N, \Pi) \\
&\quad = \sum_{i=1}^{N} H(x_i, y_i) + H(\Pi) - (H(\Pi) + \sum_{i=1}^{N} H(x_i, y_i \mid x_1, \ldots, x_{i-1}, y_1, \ldots, y_{i-1}, \Pi)) \\
&\quad = \sum_{i=1}^{N} (H(x_i, y_i) - H(x_i, y_i \mid x_1, \ldots, x_{i-1}, y_1, \ldots, y_{i-1}, \Pi)) \\
&\quad \geq \sum_{i=1}^{N} (H(x_i, y_i) - H(x_i, y_i \mid \Pi)) \\
&\quad = \sum_{i=1}^{N} I(x_i, y_i; \Pi) \ ,
\end{aligned}$$

where $H(x_i, y_i \mid x_1, \ldots, x_{i-1}, y_1, \ldots, y_{i-1}, \Pi) \leq H(x_i, y_i \mid \Pi)$ holds, since x_i and y_i's are independent, but Π may still depend on them. □

Below we restate Lemma 4 from Section 4 and give a proof.

Lemma 4 (Information to Hellinger). *Let X and Y be random variables and $Y_x := Y|X = x$, i.e., Y conditioned on $X = x$. Then*

$$\mathbb{E}_{x \in X}[\Delta^2_{\mathsf{Hel}}(Y, Y_x)] \leq I(X; Y) \ .$$

Proof. Let $P(y|x) := \Pr[Y = y|X = x]$, $P(x) := \Pr[X = x]$, and $P(y) := \Pr[Y = y]$. Then

$$\begin{aligned}
\mathbb{E}_{x \in X}[\Delta^2_{\mathsf{Hel}}(Y, Y_x)] &= \mathbb{E}_{x \in X} \left(1 - \sum_{y \in Y} \sqrt{P(y|x) \cdot P(y)}\right) \\
&= \sum_{x \in X} P(x) \cdot \left(1 - \sum_{y \in Y} \sqrt{P(y|x) \cdot P(y)}\right) \\
&= \sum_{x \in X} P(x) - \sum_{x, y \in X \times Y} P(x) \sqrt{P(y|x) \cdot P(y)}\big)
\end{aligned}$$

$$
\begin{aligned}
&= 1 - \sum_{x,y \in X \times Y} \sqrt{P(y|x) \cdot P(x) \cdot P(x) \cdot P(y)} \\
&= 1 - \sum_{x,y \in X \times Y} \sqrt{P(y,x) \cdot P(x) \cdot P(y)} \\
&= \Delta^2_{\mathsf{Hel}}(P(x,y), P(x) \cdot P(y)) \\
&\leq D_{\mathrm{KL}}(P(x,y) \parallel P(x) \cdot P(y)) \qquad (1) \\
&= I(X;Y) \qquad (2)
\end{aligned}
$$

Line (1) follows from Lemma 16 and Line (2) follows from the fact that $I(X;Y) = D_{\mathrm{KL}}(P(x,y) \parallel P(x)P(y))$. □

Lemma 16. *For any random variable $z \in D$ and probability distributions $P(z)$ and $Q(z)$ it holds that*

$$
D_{\mathrm{KL}}(P(z) \parallel Q(z)) \geq \Delta^2_{\mathsf{Hel}}(P(z), Q(z)) \ .
$$

Proof. We have

$$
\begin{aligned}
\Delta^2_{\mathsf{Hel}}(P(z), Q(z)) &= 1 - \sum_{z \in D} \sqrt{P(z) \cdot Q(z)} \\
&\leq -\log \sum_{z \in D} \sqrt{P(z) \cdot Q(z)} \\
&\leq -\log \sum_{z \in D} \big(P(z) \cdot Q(z)\big) \\
&= -\log \sum_{z \in D} \big(P(z) \cdot \frac{Q(z)}{P(z)}\big) \\
&= -\log \mathbb{E}_{z \in D} \frac{Q(z)}{P(z)} \\
&\leq -\mathbb{E}_{z \in D} \log \frac{Q(z)}{P(z)} \\
&= -\sum_{z \in D} P(z) \cdot \log \frac{Q(z)}{P(z)} \\
&= D_{\mathrm{KL}}(P(z) \parallel Q(z))
\end{aligned}
$$

The first inequality above holds $1 - x \leq -\log x$. The last inequality is an application of Jensen, which states that $\mathbb{E}[f(X)] \leq f(\mathbb{E}(X))$ for any concave f. □

C.2 Proof of the lower bound

We are now ready to prove the lower bound for distributional communication complexity of set-disjointness. Recall Theorem 2 stated in Section 4:

Theorem 2 (Set-disjointness lower bound). *Let $N \in \mathbb{N}$ and assume $p, q \in (0, 1/2]$ with $p \leq q$ and $pq = 1/(\delta N)$ for some $\delta > 1$. Let $\mu(p, q)$ be the product binomial distribution over subsets $S, T \subseteq \{0, 1\}^n$. Let*

$$p_0 := \Pr[\mathrm{DISJ}(S, T) = 0] = 1 - (1 - 1/(\delta N))^N = 1 - e^{-1/\delta} + \mathcal{O}(1/(2\delta^2 N))$$

and assume

$$\varepsilon < \frac{(\delta - 1)p_0}{(4 + \delta)} \ .$$

Then

$$\mathsf{D}^{\varepsilon}_{\mu(p,q)}(\mathrm{DISJ}) \geq \frac{Np}{8} \cdot \big((\delta - 1)p_0 - (4 + \delta)\varepsilon\big)^2 \ .$$

Proof. Let π be a deterministic protocol with error at most ε:

$$\Pr_{(S,T)\sim\mu}[\pi(X, Y) = \mathrm{DISJ}(S, T)] \geq 1 - \varepsilon \ ,$$

where X and Y are bit string representations of S and T. Let $\Pi(X, Y)$ denote a random variable for the transcripts of protocol π on inputs (X, Y) such that the corresponding sets are drawn from μ, i.e., $(S, T) \sim \mu$. Let $N := 2^n$ and write $X = (x_1, \ldots, x_N)$ and $Y = (y_1, \ldots, y_N)$ where $x_i, y_i \in \{0, 1\}$. For random variables A and B, let $\mathrm{supp}(A)$ denote the support of A (i.e., the set of values that have a non-zero probability of happening), $H(A)$ denote the Shannon entropy, and $I(A; B)$ denote the mutual information between A and B.

$$\begin{aligned}
\mathsf{D}^{\varepsilon}_{\mu(p,q)}(\mathrm{DISJ}) &\geq \log |\mathrm{supp}(\Pi(X, Y))| \\
&\geq H(\Pi(X, Y)) \\
&= H(\Pi(X, Y)) + H(X, Y) - H(X, Y, \Pi(X, Y)) \\
&= I(X, Y; \Pi) \\
&= I(x_1, \ldots, x_N, y_1, \ldots, y_N; \Pi) \\
&\geq \sum_{i=1}^{N} I(x_i, y_i; \Pi) \ ,
\end{aligned}$$

where the last inequality holds by Lemma 14. Let $\Pi^i_{a,b}$ be Π conditioned on the i-th coordinates of X and Y being fixed to a and b respectively:

$$\Pi^i_{a,b}(X, Y) := \Pi(X, Y) \mid x_i = a \wedge y_i = b \ .$$

By Lemma 4

$$I(x_i, y_i; \Pi) \geq \mathbb{E}_{(a,b)}[\Delta^2_{\mathsf{Hel}}(\Pi^i_{a,b}, \Pi)] \ ,$$

where $(a,b) \sim \mathrm{Ber}(p) \times \mathrm{Ber}(q)$. Since $q \geq p$ we have that $q(1-p) \geq p(1-q)$. Since $q \leq 1/2$, we also have that $p(1-q) \geq p/2$. Thus

$$\begin{aligned}
I(x_i, y_i; \Pi) &\geq p(1-q) \cdot \Delta^2_{\mathsf{Hel}}(\Pi^i_{1,0}, \Pi) + q(1-p) \cdot \Delta^2_{\mathsf{Hel}}(\Pi^i_{0,1}, \Pi) \\
&\geq p(1-q) \cdot \left(\Delta^2_{\mathsf{Hel}}(\Pi^i_{1,0}, \Pi) + \Delta^2_{\mathsf{Hel}}(\Pi^i_{0,1}, \Pi)\right) \\
&\geq p/2 \cdot \left(\Delta^2_{\mathsf{Hel}}(\Pi^i_{1,0}, \Pi) + \Delta^2_{\mathsf{Hel}}(\Pi^i_{0,1}, \Pi)\right) \\
&\geq p/4 \cdot \left(\Delta_{\mathsf{Hel}}(\Pi^i_{1,0}, \Pi) + \Delta_{\mathsf{Hel}}(\Pi^i_{0,1}, \Pi)\right)^2 \\
&\geq p/4 \cdot \Delta^2_{\mathsf{Hel}}(\Pi^i_{1,0}, \Pi^i_{0,1}) \ .
\end{aligned}$$

The last inequality is by the triangle inequality for the metric Δ_{Hel}, and the penultimate inequality uses $x^2 + y^2 \geq (x+y)^2/2$. Hence,

$$\begin{aligned}
\mathsf{D}^{\varepsilon}_{\mu(p,q)}(\mathrm{DISJ}) &\geq N \cdot \mathbb{E}_i[I(x_i, y_i; \Pi)] \\
&\geq Np/4 \cdot \mathbb{E}_i[\Delta^2_{\mathsf{Hel}}(\Pi^i_{1,0}, \Pi^i_{0,1})] \\
&= Np/4 \cdot \mathbb{E}_i[\Delta^2_{\mathsf{Hel}}(\Pi^i_{0,0}, \Pi^i_{1,1})] \\
&\geq Np/8 \cdot \mathbb{E}_i[\Delta^2_{\mathsf{TV}}(\Pi^i_{0,0}, \Pi^i_{1,1})] \\
&\geq Np/8 \cdot \left(\mathbb{E}_i[\Delta_{\mathsf{TV}}(\Pi^i_{0,0}, \Pi^i_{1,1})]\right)^2
\end{aligned}$$

where the equality uses the cut-and-paste lemma (Lemma 5 below) which states that $\Delta^2_{\mathsf{Hel}}(\Pi^i_{1,0}, \Pi^i_{0,1}) = \Delta^2_{\mathsf{Hel}}(\Pi^i_{0,0}, \Pi^i_{1,1})$ for any protocol π. The penultimate inequality uses $\Delta_{\mathsf{TV}}(A,B) \leq \sqrt{2}\Delta_{\mathsf{Hel}}(A,B)$, which implies $\Delta^2_{\mathsf{Hel}}(\Pi^i_{0,0}, \Pi^i_{1,1}) \geq 1/2\Delta^2_{\mathsf{TV}}(\Pi^i_{0,0}, \Pi^i_{1,1})$, and the last inequality is by Jensen. Thus it remains to lower bound $\Delta_{\mathsf{TV}}(\Pi^i_{0,0}, \Pi^i_{1,1})$.

Let $E := (\mathrm{DISJ}(S,T) = 1 \wedge x_i = y_i = 0)$. By correctness,

$$\begin{aligned}
1 - \varepsilon &\leq \Pr[\Pi(X,Y) = \mathrm{DISJ}(S,T)] \\
&= \Pr[\Pi(X,Y) = \mathrm{DISJ}(S,T) \wedge E] + \Pr[\Pi(X,Y) = \mathrm{DISJ}(S,T) \wedge \neg E] \\
&\leq \Pr[\Pi(X,Y) = \mathrm{DISJ}(S,T) \wedge E] + 1 - \Pr[E] \\
&= \Pr[\Pi(X,Y) = 1 \wedge E] + 1 - \Pr[E] \ .
\end{aligned}$$

Hence

$$\Pr[\Pi(X,Y) = 1 \wedge E] \geq \Pr[E] - \varepsilon \ .$$

For every i we have

$$\begin{aligned}
\Pr[\Pi^i_{0,0}(X,Y) = 1] &= \Pr[\Pi(X,Y) = 1 \wedge x_i = y_i = 0] / \Pr[x_i = y_i = 0] \\
&\geq \Pr[\Pi(X,Y) = 1 \wedge E] / \Pr[x_i = y_i = 0] \\
&\geq (\Pr[E] - \varepsilon) / \Pr[x_i = y_i = 0] \\
&= \Pr[\mathrm{DISJ}(S,T) = 1 | x_i = y_i = 0] - \varepsilon / \Pr[x_i = y_i = 0] \\
&\geq \Pr[\mathrm{DISJ}(S,T) = 1] - \varepsilon / \Pr[x_i = y_i = 0] \\
&\geq \Pr[\mathrm{DISJ}(S,T) = 1] - \varepsilon / ((1-p)(1-q))
\end{aligned}$$

Complementing the output bit and setting $p_0 := \Pr[\mathrm{DISJ}(S,T) = 0]$ we get

$$\Pr[\Pi^i_{0,0}(X,Y) = 0] \leq p_0 + \varepsilon/((1-p)(1-q)) \leq p_0 + 4\varepsilon \ .$$

Now we look at $x_i = y_i = 1$. In this case $\mathrm{DISJ}(S,T) = 0$. A natural strategy at this point would be to argue that π outputs 0 with good probability for every i (since it is highly correct). This, however, does not work as $\Pr[x_i = y_i = 1]$ can be small. Despite this, $\Pr[\exists i : x_i = y_i = 1]$ is high, and the protocol π over a *random* choice of i should output 0 with good probability. We proceed as follows. Note

$$1 - \varepsilon \leq \Pr[\Pi(X,Y) = 0 \wedge \mathrm{DISJ}(S,T) = 0] + (1 - p_0)$$

Thus $\Pr[\Pi(X,Y) = 0 \wedge \mathrm{DISJ}(S,T) = 0] \geq p_0 - \varepsilon$. Now

$$\begin{aligned}
\mathbb{E}_i[\Pr[\Pi^i_{1,1}(X,Y) = 0]] &= 1/N \cdot \sum \Pr[\Pi(X,Y) = 0 | x_i = y_i = 1] \\
&= 1/N \cdot \sum \Pr[\Pi(X,Y) = 0 \wedge x_i = y_i = 1] / \Pr[x_i = y_i = 1] \\
&= 1/(Npq) \cdot \sum \Pr[\Pi(X,Y) = 0 \wedge x_i = y_i = 1] \\
&\geq 1/(Npq) \cdot \Pr[\Pi(X,Y) = 0 \wedge \exists i : x_i = y_i = 1] \\
&= 1/(Npq) \cdot \Pr[\Pi(X,Y) = 0 \wedge \mathrm{DISJ}(S,T) = 0] \\
&\geq 1/(Npq) \cdot (p_0 - \varepsilon)
\end{aligned}$$

Putting the above two bounds together we get

$$\mathbb{E}_i[\Delta_{\mathsf{TV}}(\Pi^i_{0,0}, \Pi^i_{1,1})] \geq (p_0 - \varepsilon)/(Npq) - p_0 - 4\varepsilon \ .$$

Since $pq = 1/(\delta N)$, we get a lower bound $(\delta - 1)p_0 - (4 + \delta)\varepsilon$, which is positive for

$$\varepsilon < \frac{(\delta - 1)p_0}{(4 + \delta)}$$

Thus this proof does not extend to large error. Overall for ε satisfying the above we get

$$\mathsf{D}^{\varepsilon}_{\mu(p,q)}(\mathrm{DISJ}) \geq \frac{Np}{8} \cdot ((\delta - 1)p_0 - (4 + \delta)\varepsilon)^2 \ .$$

□

We restate Lemma 5 from Section 4 and include a proof below.

Lemma 5 (Cut-and-Paste). *Let $\Pi(X,Y)$ denote a random variable for the transcripts of a deterministic protocol on input bit strings (X,Y) such that the corresponding sets S and T are drawn from μ, i.e., $S,T \sim \mu$. Let $a,b \in \{0,1\}$ and define $\Pi^i_{a,b}(X,Y) := \Pi(X,Y) \mid x_i = a \wedge y_i = b$. Then for each i, it holds that*

$$\Delta^2_{\mathsf{Hel}}(\Pi^i_{0,0}, \Pi^i_{1,1}) = \Delta^2_{\mathsf{Hel}}(\Pi^i_{0,1}, \Pi^i_{1,0}) \ .$$

Proof. By definition of Δ_{Hel} it suffices then to show that for each i and any transcript τ,

$$\Pr[\Pi^i_{0,0} = \tau] \Pr[\Pi^i_{1,1} = \tau] = \Pr[\Pi^i_{0,1} = \tau] \Pr[\Pi^i_{1,0} = \tau] \ .$$

By the rectangle property of protocols, any transcript τ corresponds to a rectangle $R_\tau = S_\tau \times T_\tau$. Thus

$$\Pr[\Pi^i_{a,b} = \tau] = \Pr[X_a^{-i} \in S_\tau \wedge Y_b^{-i} \in T_\tau]$$

where X^{-i} (resp. Y^{-i}) is the input with i-th coordinate removed. By the independence of X and Y we can write

$$\Pr[X_a^{-i} \in S_\tau \wedge Y_b^{-i} \in T_\tau] = \Pr[X_a^{-i} \in S_\tau] \Pr[Y_b^{-i} \in T_\tau]$$

That is

$$\Pr[\Pi^i_{a,b} = \tau] = A^i_a(\tau) B^i_b(\tau).$$

for some A^i_a and B^i_b. Thus,

$$\begin{aligned} \Pr[\Pi^i_{0,0} = \tau] \Pr[\Pi^i_{1,1} = \tau] &= A^i_0(\tau) B^i_0(\tau) A^i_1(\tau) B^i_1(\tau) \\ &= A^i_0(\tau) B^i_1(\tau) A^i_1(\tau) B^i_0(\tau) \\ &= \Pr[\Pi^i_{0,1} = \tau] \Pr[\Pi^i_{1,0} = \tau] \ . \end{aligned}$$

□

C.3 Lower bound for $\delta \geq 0.8$

In our analysis above we showed that

$$\mathbb{E}_i[\Pr[\Pi^i_{1,1}(X, Y) = 0]] \geq 1/(Npq) \cdot (p_0 - \varepsilon)$$

This bound does not quite match what we would intuitively expect: when ε is zero (or is extremely close to zero) we would expect that the protocol will output 0 with probability 1, since under the condition $x_i = y_i = 1$ the two sets are not disjoint.

On top of this mismatch with intuition, this bound prevents us from choosing parameters such that the concatenation combiner is *compressing*. The stretch of the combined function is however $s := 2n + s_1 + s_2 - n = n + s_1 + s_2$. A look at the final bound obtained in Theorem 2 reveals that it is only meaningful for $\delta > 1$ and becomes void even if δ is set to 1. Since $pq = 1/(\delta N)$, $p = 1/2^{n+s_1}$ and $q = 1/2^{n+s_2}$ we must have that $n + s_1 + s_2 = \log \delta$. Hence $s = \log \delta > 0$ and the function cannot be compressing.

We prove an improved lower bound for the protocol correctness in case $x_i = y_i = 1$. To the best of our knowledge this analysis is new.

Let $Z := (X, Y)$, $Z_i := (x_i = y_i = 1)$ and $w(Z) := |\{i \mid Z_i\}|$. We assume $Npq = 1/\delta$ and compute $\mathbb{E}_i[\Pr[\Pi^i_{1,1}(X, Y) = 0]]$ as follows.

$$= \frac{1}{Npq} \sum_i \Pr[\Pi = 0 \wedge Z_i]$$

$$
\begin{aligned}
&= \frac{1}{Npq} \sum_i \Pr[\Pi = 0] \Pr[Z_i | \Pi = 0] \\
&= \delta \Pr[\Pi = 0] \sum_i \sum_k \Pr[Z_i | \Pi = 0] \\
&= \delta \Pr[\Pi = 0] \sum_i \sum_k \Pr[Z_i \wedge w(Z) = k | \Pi = 0] \\
&= \delta \Pr[\Pi = 0] \sum_{k>0} \sum_i \Pr[Z_i \wedge w(Z) = k | \Pi = 0] \\
&= \delta \Pr[\Pi = 0] \sum_{k>0} \sum_i \sum_{z, w(z)=k} \Pr[Z_i \wedge Z = z | \Pi = 0] \\
&= \delta \Pr[\Pi = 0] \sum_{k>0} \sum_{z, w(z)=k} \sum_i \Pr[Z = z | \Pi = 0] \Pr[Z_i | \Pi = 0 \wedge Z = z] \\
&= \delta \Pr[\Pi = 0] \sum_{k>0} \sum_{z, w(z)=k} \Pr[Z = z | \Pi = 0] \sum_i \Pr[Z_i | \Pi = 0 \wedge Z = z] \\
&= \delta \Pr[\Pi = 0] \sum_{k>0} \sum_{z, w(z)=k} \Pr[Z = z | \Pi = 0] w(z) \\
&= \delta \Pr[\Pi = 0] \sum_{k>0} k \sum_{z, w(z)=k} \Pr[Z = z | \Pi = 0] \\
&= \delta \Pr[\Pi = 0] \sum_{k>0} k \Pr[w(Z) = k | \Pi = 0] \\
&= \delta \sum_{k>0} k \Pr[w(Z) = k \wedge \Pi = 0]
\end{aligned}
$$

Let $p_k := \Pr[w(Z) = k]$ and $\varepsilon_k := \Pr[\Pi = 1 | w(Z) = k]$, the probability that the protocol errs knowing that $w(Z) = k$. Then $\Pr[w(Z) = k \wedge \Pi = 0] = p_k(1 - \varepsilon_k)$. So we want to minimize the function $\sum_{k>0} k p_k (1 - \varepsilon_k)$ subject to the condition: $\sum p_k \varepsilon_k \leq \varepsilon$. This program is equivalent to

$$
\begin{aligned}
\text{Maximize:} \quad & \sum_{k>0} k p_k \varepsilon_k \\
\text{Subject to:} \quad & \sum_{k>0} p_k \varepsilon_k \leq \varepsilon \\
& 0 \leq \varepsilon_k \leq 1
\end{aligned}
$$

The solution of this problem is obtained by defining ℓ as the smallest integer such that $\sum_{k=\ell+1}^{N} p_k < \varepsilon$ and $\sum_{k=\ell}^{N} p_k \geq \varepsilon$. With this ℓ we define

$$
\varepsilon_k := \begin{cases} 0 & \text{if } 0 < k \leq \ell - 1 \ ; \\ \varepsilon - \frac{\sum_{k=\ell+1}^{N} p_k}{p_\ell} & \text{if } k = \ell \ ; \\ 1 & \text{if } k \geq \ell + 1 \ . \end{cases}
$$

We note that p_k is decreasing and has limit $\frac{1}{ek!}$. Let $\varepsilon = 1/100$. For $N \geq 8$ we have

$$\sum_{k=4}^{N} \frac{1}{ek!} \geq \varepsilon \qquad \text{and} \qquad \sum_{k=5}^{N} \frac{1}{ek!} < \varepsilon \ .$$

Hence we can set $\ell = 4$. We deduce that

$$\delta \sum_{k>0} k \Pr[w(Z) = k \wedge \Pi = 0] \geq \delta \sum_{k>0}^{\ell-1} \frac{k}{ek!} > 0.91\delta \ .$$

We get statistical distance

$$0.91\delta - (1 - (1 - 1/(\delta N))^N + 4\varepsilon) \ ,$$

Which for $\delta \geq 0.8$ is at least 0.01. This then leads to an overall lower bound of $10^{-5}Np$.

D Attacks

In this section we (informally) present attacks against the concatenation and cascade combiners. The xor combiner seems to be secure independently of its domain and co-domain sizes. The success of attacks on the other two combiners, however, depends on the sizes of the domain and co-domain of the available hash functions. For constants $\alpha > 0$ and β we call a hash function $\mathsf{H} \in \mathrm{Fun}[n, \alpha \cdot n + \beta]$ compressing if $\alpha < 1$, approximately length preserving if $\alpha = 1$, and expanding if $\alpha > 1$.

D.1 Concatenation

Consider two hash functions $\mathsf{H}_1 \in \mathrm{Fun}[n, \alpha_1 n + \beta_1]$ and $\mathsf{H}_2 \in \mathrm{Fun}[n, \alpha_2 n + \beta_2]$ used in the concatenation combiner. The combined function is in $\mathrm{Fun}[n, (\alpha_1 + \alpha_2)n + (\beta_1 + \beta_2)]$. Any point in the co-domain of H_1 will be in $\mathsf{Img}(\mathsf{H}_1)$ with probability

$$\Pr[\exists x : \mathsf{H}_1(x) = y] = 1 - \Pr[\forall x : \mathsf{H}_1(x) \neq y] = 1 - (1 - 1/2^{\alpha_1 n + \beta_1})^{2^n} \ .$$

For $\alpha_1 < 1$ this probability is negligible, for $\alpha_1 = 1$ it is constant, and for $\alpha_1 > 1$ it is overwhelming. A similar analysis holds for H_2.

We can attempt to invert $y_1^* | y_2^* = \mathsf{H}_1(x) | \mathsf{H}_2(x)$ by using the backdoor oracle to compute random preimages for y_1^* under H_1 and random preimages for y_2^* under H_2. The two sets of preimages obtained must have a non-empty intersection in order to find a valid preimage for the combined function. When either of the hash functions is expanding or approximately length preserving, the corresponding set of preimages will be of polynomial size and a preimage can be found with good probability after polynomially many inversion queries.

To attack PRG security note that whenever we do not (even) have weak OW security, we will not have weak PRG security either. However, a slightly

expanding function (say by 1 bit) can be weakly OW, and also weakly PRG secure.

Collisions for H_1 or H_2 do not necessarily lead to collisions for the combined function. Indeed, we are not able to give attacks for any stretch pattern here.

D.2 Cascade

Consider two hash functions $\mathsf{H}_1 \in \mathrm{Fun}[n, \alpha_1 n+\beta_1]$ and $\mathsf{H}_2 \in \mathrm{Fun}[\alpha_1 n+\beta_1, \alpha_2 n+\beta_2]$ used in the cascade combiner. The combined function is in $\mathrm{Fun}[n, \alpha_2 n + \beta_2]$.

When H_2 is somewhat expanding or approximately length preserving (that is, $\alpha_2 \geq \alpha_1$) we can attempt to invert the challenge $y^* := \mathsf{H}_2(\mathsf{H}_1(x^*))$ under H_2 and with non-negligible probability obtain a point z such that $z = z^*$ where $z^* := \mathsf{H}_1(x^*)$. We can then try to invert z^* under H_1 to obtain a preimage. Note that this last step will succeed if either we are lucky and get $z = z^*$ in the first step, or if H_1 is either approximately length preserving or somewhat compressing ($\alpha_1 \leq 1$) so that the obtained z will invert with good probability. This leaves the composition of a somewhat expanding H_1 with a somewhat compressing H_2 (the expand-then-compress construction) as the only option which can offer strong one-way security.

For PRG security, as seen above, if the outer function H_2 is not sufficiently compressing, a random point will fail to invert under H_2 with non-negligible probability, whereas an honestly computed PRG value always will.

For collision resistance, any collision for H_1 is also a collision for the combiner. Collisions can be easily found for H_1 in the BRO model, and hence H_1 must be injective (even for *weak* collision resistance). The probability that a random $\mathsf{H}_1 \in \mathrm{Fun}[n, m]$ with $m > n$ is injective is

$$\Pr[\mathsf{H} \text{ is injective}] = \frac{1}{(2^m)^{2^n}} \frac{2^m!}{(2^m - 2^n)!} \approx e^{-2^{2n-m}} ,$$

using Stirling's approximations. If $\alpha_1 > 2$ the exponent $2n - m$ tends to $-\infty$ as n grows, and the above probability is overwhelming. Collisions for H_2, on the other hand, do not necessarily give rise to those for the combined function. We can thus set $\alpha_1 = 3$, $\beta_1 = 0$, $\alpha_2 = 1$ and $\beta_2 = -1$ to get a combined function that compresses by 1 bit and is conjecturally collision resistant.

E Collision Resistance of the Cascade Combiner

In this section, we prove the collision resistance of the cascade combiner based on the hardness of the multi-instance 2INT problem as defined in Section 4.

Theorem 4. *Let* $\mathsf{H}_1 \in \mathrm{Fun}[n, n + s_1]$ *and* $\mathsf{H}_2 \in \mathrm{Fun}[n + s_1, n + s_1 + s_2]$ *and* $\mathsf{C}^{\mathsf{H}_1,\mathsf{H}_2}(x) := \mathsf{H}_2(\mathsf{H}_1(x))$. *Then for any adversary* $\mathcal{A}$ *against the collision resistance of* $\mathsf{C}^{\mathsf{H}_i}$ *in the 2-BRO model there is a 2-party protocol* π *against multi-instance set-intersection problem over the distribution* $\mu(p, q)$ *with* $p := 1/2^{s_1}$ *and*

$q := 2n \ln 2/2^{n+s_1+s_2}$ *and where Alice holds* 1 *set and Bob holds* $M_2 := 2^{n+s_1+s_2}$ *many sets such that*

$$\mathsf{Adv}^{\mathrm{cr}}_{\mathsf{C}^{\mathsf{H}_1,\mathsf{H}_2}}(\mathcal{A}) \le \mathsf{Adv}^{\text{mi-2int}}_{\mu}(\pi) + \sqrt{n}2^{-n/2}(1 + 2^{-s_2-s_1}) + 2^{-n}$$
$$\textit{and} \quad \mathsf{D}_{\mu}(\pi) \le \mathsf{Q}(\mathcal{A}) + 2n + (n + s_1 + s_2) \ .$$

Proof. Let $\mathcal{A}$ be an adversary that can find collisions under $\mathsf{C}^{\mathsf{H}_i}$, i.e., find two values $x, x' \in \{0,1\}^n$ such that $x \neq x'$ and $\mathsf{H}_2(\mathsf{H}_1(x)) = \mathsf{H}_2(\mathsf{H}_1(x'))$. Given such values we want to compute $y := \mathsf{H}_1(x)$ and $y' := \mathsf{H}_1(x')$ such that $y, y' \in I \cap T_i$ for some $i \in \{0,1\}^{n+s_1+s_2}$, where $T_i := \mathsf{H}_2^{-}(i)$, and $I := \mathsf{H}_1(\{0,1\}^n)$. The sets $T_i \subseteq \{0,1\}^{n+s_1}$ are Bernoulli with parameter $\Pr[y \in T_i] = 1/2^{n+s_1+s_2}$. For the reduction we need to deal with two issues. On the one hand such $I \subseteq \{0,1\}^{n+s_1}$ is, as we argued in the proof of Theorem 7, not Bernoulli, and on the other hand, Bob's sets T_i are partitioned and are not independent. That is they do not intersect and their union forms the entire $\{0,1\}^{n+s_1}$.

Suppose now that Alice holds a set S which is Bernoulli with parameter $1/2^{s_1}$ and Bob holds $2^{n+s_1+s_2}$ many independent sets Q_i that are Bernoulli in $2n \ln 2/2^{n+s_1+s_2}$. Alice and Bob need to prepare their inputs to be able to use $\mathcal{A}$ in finding an intersection in $S \cap Q_i$ for some $i \in \{0,1\}^{n+s_1+s_2}$. More precisely, Alice uses the algorithm $\mathsf{HashSam}$ from Figure 5 to simulate a hash function H_1 which has image I, and Bob adapts his sets using the algorithm ReDist from Figure 4 such they form a partition of the domain of H_2.

As shown for one-wayness of the cascade combiner in the proof of Theorem 7, a hash function returned by $\mathsf{HashSam}$ is distributed as a random oracle in $\mathrm{Fun}[n, n+s_1]$ if the starting set S is Bernoulli with parameter $1/2^{s_1}$.

Analogous to Lemma 6 we can argue that the hardness of the standard multi-instance set intersection problem, where one party holds one set and the other $M = 2^{n+s_1+s_2}$ sets, on distribution $\tilde{\mu}(\tilde{p}, \tilde{q})$ implies the hardness of the partitioned problem on the distribution $\mu(\tilde{p}, 2n \ln 2 \cdot \tilde{q})$. Let $p = \tilde{p} = 1/2^{s_1}$, $\tilde{q} = 1/2^{n+s_1+s_2}$. A similar analysis to the proof of this lemma yields that Bob's partitions Q_i are Bernoulli with parameter $2n \ln 2 \cdot \tilde{q}$. Hence Bob can simulate a random oracle H_2 using its tweaked inputs by setting $\mathsf{H}_2^{-}(i) = Q_i$ for all $i \in \{0,1\}^{n+s_1+s_2}$.

Using the simulated H_1 and H_2 Alice and Bob run $\mathcal{A}$ in tandem, answering its oracle queries using their full knowledge of their own hash functions. Once $\mathcal{A}$ terminates with some $x, x' \in \{0,1\}^n$, Alice computes two elements in the intersection as $y := \mathsf{H}_1(x)$ and $y' := \mathsf{H}_1(x')$. □