Abstract Channels and Their Robust Information-Leakage Ordering

Annabelle McIver[1], Carroll Morgan[2,⋆], Geoffrey Smith[3], Barbara Espinoza[3], and Larissa Meinicke[4]

[1] Macquarie University
annabelle.mciver@mq.edu.au
[2] University of New South Wales and NICTA
carrollm@cse.unsw.edu.au
[3] Florida International University
{smithg,bespi009}@cis.fiu.edu
[4] University of Queensland
l.meinicke@uq.edu.au

Abstract. The observable output of a probabilistic system that processes a secret input might reveal some information about that input. The system can be modelled as an information-theoretic channel that specifies the probability of each output, given each input. Given a prior distribution on those inputs, entropy-like measures can then quantify the amount of information leakage caused by the channel. But it turns out that the conventional channel representation, as a matrix, contains structure that is redundant with respect to that leakage, such as the labeling of columns, and columns that are scalar multiples of each other. We therefore introduce *abstract channels* by quotienting over those redundancies.

A fundamental question for channels is whether one is worse than another, from a leakage point of view. But it is difficult to answer this question robustly, given the multitude of possible prior distributions and leakage measures. Indeed, there is growing recognition that different leakage measures are appropriate in different circumstances, leading to the recently proposed g-leakage measures, which use gain functions g to model the operational scenario in which a channel operates: the *strong g-leakage pre-order* requires that channel A never leak more than channel B, for any prior and any gain function. Here we show that, on abstract channels, the strong g-leakage pre-order is *antisymmetric*, and therefore a *partial order*.

It was previously shown [1] that the strong g-leakage ordering is implied by a structural ordering called *composition refinement*, which requires that $A = BR$, for some channel R; but the converse was not established in full generality, left open as the so-called *Coriaceous Conjecture*. Using ideas from [2], we here confirm the *Coriaceous Conjecture*. Hence the strong g-leakage ordering and composition refinement coincide, giving our partial order both structural- and leakage-testing significance.

⋆ NICTA is funded by the Australian Government through the Department of Communications and the Australian Research Council through the ICT Centre of Excellence Program.

M. Abadi and S. Kremer (Eds.): POST 2014, LNCS 8414, pp. 83–102, 2014.

1 Introduction

A fundamental goal in computer security is the protection of confidential information from improper disclosure. Yet this goal often cannot be achieved perfectly, because certain leaks of confidential information are unavoidable. The importance of *quantitative information flow* is therefore that it enables us to say that certain information leaks are "small" and hence tolerable.

Consider a channel C that takes as input a secret X with *prior* probability distribution π, and produces (perhaps probabilistically) an observable output Y. If an adversary knows π and C, then its *initial uncertainty* about X will depend on π. But each separate output value y then allows it to update its knowledge about X's prior π to a *posterior* distribution $p_{X|y}$ via Bayesian reasoning. Hence its expected *remaining uncertainty* about X, after seeing the output of C, will depend on the set of possible posterior distributions on X and their probabilities. The *leakage* is the difference between the initial and final uncertainties.

This general quantitative framework is clear enough; but there is of course more than one way to measure the "uncertainty" associated with a probability distribution: popular choices include Shannon entropy [3], guessing entropy [4], min-entropy [5], and the family of g-entropies [1] each determined by its own gain function g. Each of those leakage measures has its own operational significance, which might or might not suit the operational scenario. Moreover, the leakage caused by some C will also depend on its prior π. As a result, if we consider the *leakage ordering* of two channels A and B (both taking X as input), it is difficult to give an answer that is *robust*, i.e. that does not depend on the particular prior and leakage measure. But such a robust ordering is indispensable if we aim to develop software through stepwise refinement, based on general laws that hold in *all contexts*.

There is such a robust order for *deterministic channels*, provided by the *Lattice of Information* [6]. Any deterministic channel from X to Y induces a *partition* on $\mathcal{X}$, where x_1 and x_2 belong to the same block iff they map to the same output.[1] That is, each block of the partition is the pre-image of some output y.

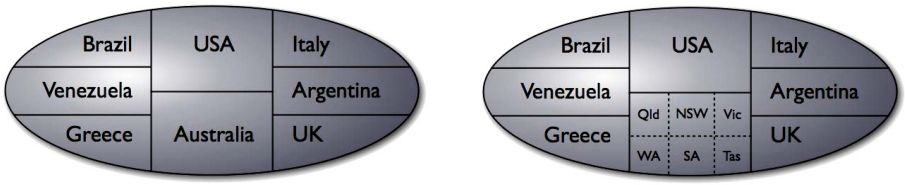
Definition 1 (Partition refinement). *Two deterministic channels A, B on input $\mathcal{X}$ are said to be in the* partition refinement *relation, written $A \sqsubseteq B$, just when the partition induced by A on $\mathcal{X}$ is refined (as a partition) by the partition induced by B: the blocks of B are formed by subdividing blocks of A.*

For example a deterministic channel A taking a secret person X to her country of birth would induce the partition in Fig. 1(a); the channel B that in some cases gives the state as well leads to Fig. 1(b).

It is intuitively clear that an adversary will *always* prefer B to A, whatever the input prior π; and this is supported by the following theorem due to Yasuoka & Terauchi, and Malacaria [7,8].

Theorem 1. *If A, B are deterministic, then $A \sqsubseteq B$ iff A never leaks more than B, on* any *prior π and under Shannon-entropy, min-entropy, or guessing-entropy leakage.*

[1] We use $\mathcal{X}$ for the set of inputs, with x being a value in $\mathcal{X}$ and X being a random variable on $\mathcal{X}$.



(a) Less refined A (b) More refined B

Fig. 1. Partition refinement

The "only if" direction of this theorem can be seen as expressing the partition refinement order's *soundness* with respect to the leakage order. More interestingly, the "if" direction can be seen as expressing its *completeness*, for it says that the *only* way for A to never leak more than B is for A's partition to be refined by B's.[2] Another way of understanding this result is to say that partition refinement is an order on deterministic channels with both a *structural-* and a *leakage-testing* characterization.

The main goal of this paper is to generalize these nice properties from *deterministic* to *probabilistic* channels. A first issue, however, is that the story for deterministic channels is not quite as nice as it appears, in that partition refinement is not in fact a *partial order* on deterministic channels, but only a *pre-order*. Because distinct deterministic channels can induce the same partition on $\mathcal{X}$ (since the particular *names* of the outputs do not matter), partition refinement is not antisymmetric. While this problem is rather obvious in the case of deterministic channels, we will see that it is more subtle for probabilistic channels, and this will lead us to introduce *abstract channels* formed by quotienting away the redundant structure of classical channel matrices.

We explore the fundamental properties of abstract channels, including their canonical representation by reduced matrices and by hyper-distributions. Turning to their robust leakage ordering, we consider a generalization of partition refinement called *composition refinement* ($\sqsubseteq_\circ$) [2,1], where $A \sqsubseteq_\circ B$ holds if A can be expressed as B followed by "post-processing". In our first major result, we show that composition refinement is antisymmetric, and therefore a partial order, on abstract channels. Next we consider the soundness and completeness of composition refinement with respect to leakage orders. It was proved in [1,9] that composition refinement implies the *strong g-leakage ordering* ($\leq_{\mathcal{G}}$), where $A \leq_{\mathcal{G}} B$ holds if A never leaks more than B, on any prior distribution and any gain function. The converse, however, was not proved in full generality, and was left as the *Coriaceous Conjecture*. In our second major result, we use ideas from [2] to prove the Coriaceous Conjecture. Hence composition refinement and the strong g-leakage ordering *coincide*, giving us a partial order on abstract channels that has both structural- and leakage-testing significance.

[2] The "if" direction is actually easy to see—for if A's partition is *not* refined by B's, then there must exist x_1 and x_2 that belong to the *same* block of B, but to *different* blocks of A. On a prior that gives non-zero probability only to x_1 and x_2, B leaks nothing about X, while A leaks everything.

In summary, our principal contributions are (1) the concept of *abstract channels*, which we argue to be the fundamental mathematical space for information-theoretic leakage; (2) the proof that composition refinement is a *partial order* on this space; and (3) the proof that composition refinement is *complete* with respect to the strong g-leakage ordering.

The rest of the paper is structured as follows: Section 2 presents preliminaries; Section 3 introduces abstract channels; Section 4 presents composition refinement and proves that it is a partial order on abstract channels; Section 5 proves that composition refinement implies the strong g-leakage ordering; Section 6 proves the converse, resolving the Coriaceous Conjecture; Section 7 gives a monadic presentation of composition refinement; Section 8 discusses limits of the information-theoretic perspective with respect to computationally-bounded adversaries; Section 9 discusses related work; and Section 10 concludes.

2 Preliminaries: Channels and Leakage Measures

We begin by recalling the basic definitions of information-theoretic channels [10]. A *channel* is a triple $(\mathcal{X}, \mathcal{Y}, C)$, where $\mathcal{X}$ and $\mathcal{Y}$ are finite sets (of secret input values and observable output values) and C is an $|\mathcal{X}|\times|\mathcal{Y}|$ *channel matrix* whose entries are between 0 and 1 and whose rows each sum to 1; the intent is that $C_{x,y}$ is the conditional probability of output y given input x. Channel C is *deterministic* if each entry of C is either 0 or 1, implying that each input row contains a single 1 which identifies its unique corresponding output.

For *prior distribution* π on $\mathcal{X}$, the *joint distribution* on $\mathcal{X}\times\mathcal{Y}$ is $p(x, y) = \pi[x]C_{x,y}$, with jointly distributed random variables X, Y whose marginal probabilities are given by $p(x) = \sum_y p(x, y)$ and $p(y) = \sum_x p(x, y)$, and whose conditional probabilities are given by $p(y|x) = {}^{p(x,y)}/_{p(x)}$ (if $p(x)$ is non-zero) and $p(x|y) = {}^{p(x,y)}/_{p(y)}$ (if $p(y)$ is non-zero). Note that p_{XY} is the *unique* joint distribution that recovers π and C, in that $p(x) = \pi[x]$ and $p(y|x) = C_{x,y}$ (if $p(x)$ is non-zero).[3]

For a given y (such that $p(y)$ is non-zero), the conditional probabilities $p(x|y)$ for each $x\in\mathcal{X}$ form the *posterior distribution* $p_{X|y}$, which is the knowledge that the adversary learns about X by seeing output y.

Example 1. Given $\mathcal{X} = \{x_1, x_2, x_3\}$, and $\mathcal{Y} = \{y_1, y_2, y_3, y_4\}$, and (the uniform) prior $\pi = (1/3, 1/3, 1/3)$, consider channel C and its associated joint matrix J as follows:

C	y_1	y_2	y_3	y_4
x_1	1	0	0	0
x_2	0	1/2	1/4	1/4
x_3	1/2	1/3	1/6	0

leads via π to the joint matrix

J	y_1	y_2	y_3	y_4
x_1	1/3	0	0	0
x_2	0	1/6	1/12	1/12
x_3	1/6	1/9	1/18	0

.

By summing J's columns we get the (marginal) distribution $p_Y = (1/2, 5/18, 5/36, 1/12)$ and by normalizing the columns we get the posterior distributions $p_{X|y_1} = (2/3, 0, 1/3)$, $p_{X|y_2} = (0, 3/5, 2/5)$, $p_{X|y_3} = (0, 3/5, 2/5)$ and $p_{X|y_4} = (0, 1, 0)$. □

[3] When necessary to avoid ambiguity, we write distributions with subscripts, e.g. p_{XY} or p_Y.

Leakage measures are defined based on various entropy-like measures of the prior distribution π and the posterior distributions $p_{X|y}$, together with their probabilities $p(y)$.

Shannon leakage is based on the Shannon entropy [3] of the prior distribution, $H(\pi) = -\sum_x \pi[x] \log \pi[x]$, and the expected Shannon entropy of the posterior distributions, $H(\pi, C) = \sum_y p(y)H(p_{X|y})$. The Shannon leakage is the difference $H(\pi) - H(\pi, C)$, which is equal to the mutual information $I(\pi, C)$.[4]

Guessing entropy leakage is based on the guessing entropy [4] of the prior distribution, $G(\pi) = \sum_i i\,\pi[x_i]$, with $\mathcal{X}$ indexed in non-increasing probability order, and on the expected guessing entropy of the posterior distributions $G(\pi, C) = \sum_y p(y)G(p_{X|y})$. The guessing entropy leakage is the difference $G(\pi) - G(\pi, C)$.

The operational significance of both Shannon entropy and guessing entropy can be stated in terms of the expected number of brute-force guesses that the adversary would need to find the secret.[5] But this is not really satisfactory for confidentiality, because the expected number of brute-force guesses needed to find the secret can be high even if the adversary has a high probability of guessing the secret successfully in just one try. For this reason we consider *min-entropy leakage* [5], which is based on the prior *vulnerability* of the secret to be guessed in one try $V(\pi) = \max_x \pi[x]$, and on the expected vulnerability of the posterior distributions $V(\pi, C) = \sum_y p(y)V(p_{X|y})$. The prior- and posterior min-entropies are obtained by taking the negative logarithm of the vulnerability: $H_\infty(\pi) = -\log V(\pi)$ and $H_\infty(\pi, C) = -\log V(\pi, C)$. The min-entropy leakage $\mathcal{L}(\pi, C)$ is the difference $H_\infty(\pi) - H_\infty(\pi, C)$ or, equivalently, the logarithm of the ratio of the posterior- and prior vulnerabilities, that is $\log {}^{V(\pi,C)}/_{V(\pi)}$.

While vulnerability is clearly important for confidentiality, it implicitly assumes an operational scenario in which the adversary gains only by guessing the secret *exactly*, and in *one try*. For this reason, *g-leakage* [1] generalizes vulnerability to incorporate a *gain function* g, the choice of which allows the modelling of differing operational scenarios. In each scenario, there will be some set $\mathcal{W}$ of *guesses* that the adversary could make about the secret, and for any guess w and secret value x, there will be some *gain* $g(w, x)$ that the adversary gets by having chosen w when the secret's actual value was x; gains are assumed to range from 0 (when w has no value at all) to 1 (when w is ideal). Formally, $g\colon \mathcal{W}\times\mathcal{X} \to [0, 1]$, where $\mathcal{W}$ is a finite, non-empty set. Given a gain function g, the prior *g-vulnerability* is defined as the maximum expected gain over all possible guesses: that is $V_g(\pi) = \max_w \sum_x \pi[x]g(w, x)$. The posterior g-vulnerability, the g-entropy and the g-leakage are then defined as for min-entropy leakage: we have $V_g(\pi, C) = \sum_y p(y)V_g(p_{X|y})$, and $H_g(\pi) = -\log V_g(\pi)$, and $H_g(\pi, C) = -\log V(\pi, C)$ and $\mathcal{L}_g(\pi, C) = H_g(\pi) - H_g(\pi, C) = \log {}^{V_g(\pi,C)}/_{V_g(\pi)}$.

In particular, a gain function g that gives gain 1 for guessing the secret correctly and 0 otherwise makes g-leakage coincide with min-entropy leakage: it is thus a special case. But gain functions can do much more. As explained in [1], they can model a wide variety of practical operational scenarios, including those where the adversary benefits from guessing a value *close* to the secret, guessing a *part* of the secret, guessing a *property* of the secret or guessing the secret within some bounded number of tries. They can also model scenarios where there is a *penalty* for incorrect guesses.

[4] The more usual notation for these quantities is $H(X)$, $H(X|Y)$, and $I(X; Y)$.

[5] For Shannon entropy, this follows from a result by Massey [4].

3 Abstract Channels Capture the Essence of Leakage

For a fixed channel and prior, it can easily happen that *distinct* output values y, y' in $\mathcal{Y}$ give rise to the *same* posterior distribution on $\mathcal{X}$. In that case there is actually no benefit to the adversary from distinguishing outputs y, y', since each gives the same knowledge about X. Furthermore, the output *values* y make no difference either: all that matters for any output y is its associated posterior distribution $p_{X|y}$. This implies that the result of a channel, as far as leakage is concerned, should simply be a *distribution* on posterior distributions; following [2] we call this a *hyper-distribution*.

Example 2. Returning to channel C from Ex. 1, we notice that its outputs $y_{2,3}$ produce the same posterior distribution, i.e. that $p_{X|y_2} = p_{X|y_3}$. Hence the hyper-distribution produced by C on π has only three columns rather than four:[6]

C	$1/2$	$15/36$	$1/12$
x_1	$2/3$	0	0
x_2	0	$3/5$	1
x_3	$1/3$	$2/5$	0

In this representation the columns are normalised, and are labelled by their associated marginal probabilities: the $\mathcal{Y}$-values have been removed. Note that the probability $15/36$ of the middle posterior distribution is found by adding $p(y_2) + p(y_3)$, that is $5/18 + 5/36$.

□

We capture these two abstractions in the following definition:

Definition 2 (Abstract channel). *The leakage semantics of a channel matrix is the mapping that it gives from priors to hyper-distributions.*

We call such a mapping an abstract channel.

The following theorem reassures us that we have not abstracted too much.

Theorem 2. *The usual leakage measures are well defined on abstract channels.*

Proof. As we saw in §2, under min-entropy leakage vulnerability is $V(\pi) = \max_x \pi[x]$, and posterior vulnerability is $V(\pi, C) = \sum_y p(y)V(p_{X|y})$. Hence the column *labels* y make no difference. Moreover, if $p_{X|y} = p_{X|y'}$ then the posterior vulnerability is unaffected by merging outputs y and y', since then

$$p(y)V(p_{X|y}) + p(y')V(p_{X|y'}) \quad = \quad p(y \vee y')V(p_{X|y}).$$

Other leakage measures, such as Shannon-based mutual information, behave similarly. □

Taking this abstracted, semantic viewpoint makes us realise that the conventional, channel-matrix representation can contain *redundant information* as far as leakage is concerned, namely (1) *labels* on columns, (2) columns that are *all zero*, representing outputs that can never occur, and (3) *similar* columns, which are columns that are scalar multiples of each other and therefore yield the same posterior distributions.[7] By eliminating this redundant information, we obtain a well defined *reduced matrix*:

[6] The block representation of a hyper-distribution has probabilities in its top row, rather than $\mathcal{Y}$-values.

[7] These can be seen as analogous to redundant information in computer programs, like the names of local variables, dead code, and if-statements with identical branches. Case (2) could be seen as an instance of Case (3) with a scaling factor of zero; but then similarity would not be symmetric.

Definition 3. *The reduced matrix C^r of a channel matrix C is formed by deleting output labels and all-zero columns, then adding similar columns together, and finally ordering the resulting columns lexicographically.*

Theorem 3. *Any channel matrix C has the same leakage semantics as its reduction C^r.*

Proof. Output labels, all-zero columns, and column ordering all have no effect on the hyper-distribution. And similar columns each contribute weight to the same posterior distribution; hence merging them leaves the hyper-distribution unchanged. □

A reduced matrix hence serves as a *canonical representation* of an abstract channel.

Corollary 1. *Channels C, D represent the same abstract channel just when $C^r = D^r$.*

Example 3. Given $\mathcal{X} = \{x_1, x_2, x_3\}$ we consider the following two channels C, D:

C	y_1	y_2	y_3
x_1	1	0	0
x_2	1/4	1/2	1/4
x_3	1/2	1/3	1/6

D	z_1	z_2	z_3
x_1	2/5	0	3/5
x_2	1/10	3/4	3/20
x_3	1/5	1/2	3/10

.

These channels *as matrices* are different — but *as abstract channels* they are the same. Indeed both map prior distribution $\pi = (p_1, p_2, p_3)$ to the same hyper-distribution:

	$(4p_1 + p_2 + 2p_3)/4$	$(3p_2 + 2p_3)/4$
x_1	$\frac{4p_1}{4p_1+p_2+2p_3}$	0
x_2	$\frac{p_2}{4p_1+p_2+2p_3}$	$\frac{3p_2}{3p_2+2p_3}$
x_3	$\frac{2p_3}{4p_1+p_2+2p_3}$	$\frac{2p_3}{3p_2+2p_3}$

To understand this, note that the second and third columns of C are *similar* (indeed column 2 is two times column 3). In the same way, columns 1 and 3 of D are similar (indeed column 1 is two-thirds times column 3). Hence A, B have the *same* reduced matrix, as shown here at right:

$$C^r = D^r = \begin{array}{|c|cc|}\hline x_1 & 1 & 0 \\ x_2 & 1/4 & 3/4 \\ x_3 & 1/2 & 1/2 \\ \hline\end{array}$$

□

While we have said that an abstract channel is a mapping from priors to hyper-distributions, in fact the mappings that come from channel matrices are highly constrained. Write $\lceil\pi\rceil$ for the *support* of distribution π, that is those elements (of $\mathcal{X}$) to which it assigns non-zero probability. Then we have

Theorem 4. *An abstract channel C with input $\mathcal{X}$ is completely determined by its behaviour on any* full-support *prior π, that is one with $\lceil\pi\rceil{=}\mathcal{X}$.*

Proof. If full-support π yields a certain hyper-distribution then, by scaling each of the posterior distributions with its probability, we recover the joint matrix of C^r under π. And normalizing the rows of the joint matrix gives C^r. □

It follows that we can also canonically represent an abstract channel by the hyper-distribution that it produces on (for instance) the uniform prior π_u — indeed we showed such a hyper-distribution in Ex. 2.[8]

4 Generalizing Partition Refinement to Composition Refinement

We now return our attention to the question of whether we can generalize the partition refinement $A \sqsubseteq B$ of Def. 1 in §1 from deterministic to probabilistic channels. Our criteria for success will include an investigation (in §5) of the situations in which the generalisation is *sound* in the sense that $A \sqsubseteq B$ implies that A's leakage does not exceed B's, and *complete* in that the generalisation fails only if there really is such a situation in which A leaks more than B.

In the deterministic case, A's partition is refined by B's just if we can convert from B to A by doing a "post-processing" step in which certain of B's outputs are *merged* — this corresponds to "anti-refinement" of partitions achieved by merging regions (just as federating the states of Australia takes us from Fig. 1(b) back to Fig. 1(a)). That is, we can express A as the *cascade* [12] of B and a channel R_{merge}, so that A is the *matrix product* of B and R_{merge}.[9] And, unlike partition refinement, this new formulation applies to probabilistic as well as deterministic channels.

Definition 4. *For channels A, B we say that A is* composition refined *by B, written $A \sqsubseteq_\circ B$, just when there exists a channel R such that $A = BR$.*

(Note that this definition appears in [1,9].)

On channel matrices, the composition-refinement relation is easily seen to be reflexive (since $C = CI$) and transitive (since $A = BR_1$ and $B = CR_2$ implies $A = (CR_2)R_1 = C(R_2R_1)$) — and so it is a preorder. But it is *not* antisymmetric, as can be seen from C, D in Ex. 3, where we have both $C \sqsubseteq_\circ D$ and $D \sqsubseteq_\circ C$:

C	y_1	y_2	y_3
x_1	1	0	0
x_2	$1/4$	$1/2$	$1/4$
x_3	$1/2$	$1/3$	$1/6$

$=$

D	z_1	z_2	z_3
x_1	$2/5$	0	$3/5$
x_2	$1/10$	$3/4$	$3/20$
x_3	$1/5$	$1/2$	$3/10$

R_1	y_1	y_2	y_3
z_1	1	0	0
z_2	0	$2/3$	$1/3$
z_3	1	0	0

and

D	z_1	z_2	z_3
x_1	$2/5$	0	$3/5$
x_2	$1/10$	$3/4$	$3/20$
x_3	$1/5$	$1/2$	$3/10$

$=$

C	y_1	y_2	y_3
x_1	1	0	0
x_2	$1/4$	$1/2$	$1/4$
x_3	$1/2$	$1/3$	$1/6$

R_2	z_1	z_2	z_3
y_1	$2/5$	0	$3/5$
y_2	0	1	0
y_3	0	1	0

However, if we restrict to *abstract channels*, we find that composition refinement is better behaved: it becomes a true partial order (Thm. 6 below). We now prove that fact, our first major result.

[8] In the more general setting of *Hidden Markov Models* [11], however, such functions from priors to hyper-distributions do not have the property of Thm. 4 — they are strictly more general.

[9] Indeed this equivalence was noted in Theorem 1 of [6].

Lemma 1 (Jensen's inequality for abstract channels). *Let $\mathcal{A}$ and $\mathcal{B}$ be abstract channels, with $(A, \mathcal{X}, \mathcal{Z})$ and $(B, \mathcal{X}, \mathcal{Y})$ their presentation as reduced matrices, and let F be a concave ($\frown$) function from distributions on $\mathcal{X}$ to the reals. If $A{=}BR$ for some channel matrix R then, for any full-support prior π, we have $F(\pi, A) \geq F(\pi, B)$ where as usual $F(\pi, A) = \sum_z p(z)F(p_{X|z})$.*

Furthermore, if $\mathcal{A}{\neq}\mathcal{B}$ and F is strictly concave, then the inequality is strict.

Proof. Our proof relies on Jensen's inequality [10], that if $\lambda_1, \lambda_2, \ldots \lambda_N$ are coefficients in $[0, 1]$ that sum to one, and F is concave, then $\sum_n \lambda_n F(x_n) \leq F(\sum_n \lambda_n x_n)$.

We use the following matrix notation. Given matrix M with row labels $\mathcal{X}$ and column labels $\mathcal{Y}$, we write $M_{x,y}$ to denote the (x, y) entry and $M_{-,y}$ to denote column y. A fundamental property of matrix multiplication is that $(MN)_{-,z} = M(N_{-,z})$, i.e. that column z of MN is a linear combination of the columns of M, with column z of N as the coefficients, and thus that in fact the parentheses above are not necessary.[10]

We write D_π to denote the diagonal matrix with prior π on its diagonal, so that $D_\pi A$ is the joint matrix giving p_{XZ}. Note that because A is reduced and π is full support, the columns of $D_\pi A$ are all non-zero and non-similar; hence normalizing these columns is well defined and gives the *distinct* posterior distributions $p_{X|z} = {}^1\!/_{p(z)}\, D_\pi A_{-,z}$ where $p(z)$ is the (necessarily nonzero) sum of column z. For B, similarly, the posterior distributions $p_{X|y}$ are distinct, and $p_{X|y} = {}^1\!/_{p(y)}\, D_\pi B_{-,y}$.

We now show that $F(\pi, A) \geq F(\pi, B)$ under the conditions given: first we have

$$
\begin{array}{lll}
 & F(\pi, A) & \\
= & \sum_z p(z)\, F(p_{X|z}) & \text{“defn. } F(\pi, A)\text{”} \\
= & \sum_z p(z)\, F({}^1\!/_{p(z)}\, D_\pi A_{-,z}) & \text{“} p_{X|z} = {}^1\!/_{p(z)}\, D_\pi A_{-,z}\text{”} \\
= & \sum_z p(z)\, F({}^1\!/_{p(z)}\, D_\pi BR_{-,z}) & \text{“}A{=}BR\text{”} \\
= & \sum_z p(z)\, F({}^1\!/_{p(z)}\, D_\pi\, (\sum_y B_{-,y}R_{y,z})\,) & \text{“} BR_{-,z} = \sum_y B_{-,y}R_{y,z}\text{”} \\
= & \sum_z p(z)\, F(\,\sum_y ({}^{R_{y,z}\, p(y)}\!/_{p(z)})({}^1\!/_{p(y)}\, D_\pi B_{-,y})\,) & \text{“reorganising”} \\
= & \sum_z p(z)\, F(\,\sum_y ({}^{R_{y,z}\, p(y)}\!/_{p(z)})(p_{X|y})\,) \quad , & \text{“} p_{X|y} = {}^1\!/_{p(y)}\, D_\pi B_{-,y}\text{”}
\end{array}
$$

which contains F applied to a convex combination ($\sum_y$) whose coefficients $R_{y,z}\, p(y)/p(z)$ we now show are suitable for the use of Jensen. They sum to one because

$$
\begin{array}{lll}
 & \sum_y R_{y,z}\, p(y) & \\
= & \sum_y R_{y,z}\, \sum_x (D_\pi B)_{x,y} & \text{“} p(y) = \sum_x (D_\pi B)_{x,y}\text{”} \\
= & \sum_{x,y} R_{y,z}\, (D_\pi B)_{x,y} & \text{“distributive law”} \\
= & \sum_x (D_\pi BR)_{x,z} & \text{“defn. matrix multiplication”} \\
= & \sum_x (D_\pi A)_{x,z} & \text{“}A = BR\text{”} \\
= & p(z)\,. & \text{“defn. } p(z)\text{”}
\end{array}
$$

With that done, we continue

$$
\begin{array}{lll}
\ldots = & \sum_z p(z)\, F(\,\sum_y ({}^{R_{y,z}\, p(y)}\!/_{p(z)})(p_{X|y})\,) & \text{“from above”} \\
\geq & \sum_z p(z)\, \sum_y ({}^{R_{y,z}\, p(y)}\!/_{p(z)})\, F(p_{X|y}) & \text{“}(*)\text{ Jensen wrt concave } F\text{”} \\
= & \sum_y p(y)\, F(p_{X|y}) \sum_z R_{y,z} & \text{“simplify”}
\end{array}
$$

[10] This is just associativity wrt post-multiplication by a column vector with one at row z and zeroes elsewhere.

$$
\begin{array}{ll}
=\sum_y p(y)\, F(p_{X|y}) & \text{“}\textstyle\sum_z R_{y,z} = 1\text{”} \\
=F(\pi, B) \quad , & \text{“defn. } F(\pi, B)\text{”}
\end{array}
$$

so that $F(\pi, A) \geq F(\pi, B)$ as claimed.

Now we suppose that $\mathcal{A}{\neq}\mathcal{B}$ and F is strictly concave.

A strict form of Jensen's inequality is that if $\lambda_1, \lambda_2, \ldots \lambda_N$ are coefficients in $[0, 1]$ that sum to one, with at least one $\lambda_n{\neq}1$, and F is strictly concave, and the x_n's are all distinct, then $\sum_n \lambda_n F(x_n) < F(\sum_n \lambda_n x_n)$. This will give strict inequality at $(*)$ above.

Because B is reduced, the distributions $p_{X|y}$ (the normalised columns of $D_\pi B$) are distinct; otherwise B would have similar columns. Those are the distinct x_n's for strict Jensen.

We now consider the λ_n's, showing that at least one of them is not one. No two columns of R can have a single non-zero entry in the same row, since those two columns would generate similar columns in A, contradicting A's being reduced. Thus if all columns of R have exactly one non-zero value, since those values are alone in their rows and R is a channel matrix, in fact R must be be a permutation of the identity. But that makes A a column permutation of B, impossible if A, B are reduced and distinct.

Thus channel matrix R must have some column $R_{-,\hat{z}}$ in which at least two entries are non-zero. But from $\sum_y R_{y,\hat{z}}\, p(y) = p(\hat{z})$, proved just above, plus the fact that $p(y)$ is nowhere zero, we have at least one $\hat{y}$ (in fact, two) with $R_{\hat{y},\hat{z}}\, p(\hat{y})/p(\hat{z}) \neq 1$. This $\hat{y}$ (as n) gives the $\lambda_n{\neq}1$ for that $\hat{z}$, as application of strict Jensen to that $\hat{z}$ requires.

Those facts taken all together allow us to make step $(*)$ above strict, since for all z's (the nonstrict) Jensen applies, and for $\hat{z}$ it applies strictly. □

A consequence of Lem. 1 is the following theorem, which is itself of interest.

Theorem 5 (Strict data-processing inequality). *Let $\mathcal{A}$ and $\mathcal{B}$ be abstract channels, and write $\mathcal{A} \sqsubset_\circ \mathcal{B}$ when $\mathcal{A} \sqsubseteq_\circ \mathcal{B}$ but $\mathcal{A}{\neq}\mathcal{B}$. If $\mathcal{A} \sqsubset_\circ \mathcal{B}$ then, for any full-support prior π, the mutual information leakage of $\mathcal{A}$ is strictly less than than that of $\mathcal{B}$: that is $I(\pi, \mathcal{A}) < I(\pi, \mathcal{B})$.*[11]

Proof. We appeal to the strict concavity ($\frown$) of Shannon entropy H [13, p. 85], using H for F in Lem. 1, to conclude that $H(\pi, \mathcal{A}) > H(\pi, \mathcal{B})$. Hence $I(\pi, \mathcal{A}) = H(\pi) - H(\pi, \mathcal{A}) < H(\pi) - H(\pi, \mathcal{B}) = I(\pi, \mathcal{B})$. □

A second consequence of Lem. 1 is the partial-order property we seek.

Theorem 6 (Partial order). *Composition refinement ($\sqsubseteq_\circ$) is a partial order on abstract channels.*

Proof. Since ($\sqsubseteq_\circ$) is reflexive and transitive, we need only antisymmetry. Suppose that $\mathcal{A} \sqsubseteq_\circ \mathcal{B} \sqsubseteq_\circ \mathcal{A}$ but $\mathcal{A}{\neq}\mathcal{B}$. Then in fact $\mathcal{A} \sqsubset_\circ \mathcal{B} \sqsubset_\circ \mathcal{A}$ whence, from Thm. 5, we have $I(\pi, \mathcal{A}) < I(\pi, \mathcal{B}) < I(\pi, \mathcal{A})$ for any full-support prior π — which is impossible. □

[11] To see that this theorem is indeed a strict version of the classic data-processing inequality [10], note that if $A = BR$, where A goes from $\mathcal{X}$ to $\mathcal{Z}$, B goes from $\mathcal{X}$ to $\mathcal{Y}$, and R goes from $\mathcal{Y}$ to $\mathcal{Z}$, then for any prior π we have a Markov chain $X \to Y \to Z$. The (non-strict) data-processing inequality says that in this case $I(X; Z) \leq I(X; Y)$, which in our notation is $I(\pi, A) \leq I(\pi, B)$.

We conclude this section by completing the link with reduced channels. For channels A, B write $A \approx_\circ B$ to mean $A \sqsubseteq_\circ B \sqsubseteq_\circ A$.

Lemma 2. *For any channel C (not necessarily reduced) we have $C \approx_\circ C^r$.*

Proof. The reduced form C^r of channel C is defined in Def. 3 via a series of operations: deleting all-zero columns,[12] summing (similar) columns together, and reordering columns (lexicographically). Each of those can be effected via post-multiplication with a simple channel matrix; and so their overall effect is achieved via multiplication with the (matrix) product of all those channel matrices, again a channel matrix. Hence $C^r \sqsubseteq_\circ C$.

For the reverse direction the operations are adding an all-zero column, splitting a column into several similar columns,[13] and reordering columns. Again all of these can be achieved by post-multiplication. Hence $C \sqsubseteq_\circ C^r$, and so $C \approx_\circ C^r$ as required. □

Theorem 7 (Quotienting). *The equivalence classes induced by the preorder ($\sqsubseteq_\circ$) on channels are the same as induced by the kernel of reduction ($-^r$): that is for any channels A, B we have $A \approx_\circ B$ just when $A^r = B^r$.*

Proof. If $A \approx_\circ B$ then $A^r \approx_\circ A \approx_\circ B \approx_\circ B^r$ (Lem. 2), whence $A^r \approx_\circ B^r$ by transitivity and finally $A^r = B^r$ by antisymmetry on reduced channels (Thm. 6).

If $A^r = B^r$ then $A^r \approx_\circ B^r$ (reflexivity) whence $A \approx_\circ B$ (Lem. 2 and transitivity). □

5 Composition Refinement and Leakage Orderings

In this section we address whether ($\sqsubseteq_\circ$) is a reasonable information order to impose; as mentioned at the beginning of §4, this is related to what we have called soundness and completeness. In §5.3 we briefly discuss compositionality.

5.1 Soundness of ($\sqsubseteq_\circ$)

The soundness condition for ($\sqsubseteq_\circ$) concerns the situations in which $A \sqsubseteq_\circ B$ implies that A leaks no more than B. That is, given a situation in which (limiting) leakage is important, according to some leakage measure, in what sense is it *sound* to use ($\sqsubseteq_\circ$) to reason about that system?

In fact we can argue informally that using ($\sqsubseteq_\circ$) for our reasoning ought to be sound for any reasonable situation and associated leakage measure: if $A = BR$ for some R, then an adversary should never prefer channel A to channel B, because given channel B the adversary can always *simulate* channel A by simply post-processing the output from channel B according to channel R.

[12] This is where we depend on deleting only *all-zero* columns to proceed from C to C^r: although post-multiplication with a channel matrix can add an all-zero column, it cannot delete a column unless that column is all zero.

[13] This is where we depend on summing only *similar* columns to proceed from C to C^r: although post-multiplication with a channel matrix can sum any two columns, similar or not, it cannot in general decompose a column into a sum of *dissimilar* columns.

And indeed this property does hold for Shannon-entropy leakage, min-entropy leakage, and g-leakage. It is a generalized *data-processing inequality*, proved here[14] for the case of g-leakage.[15]

Theorem 8. *If $A \sqsubseteq_\circ B$ then the g-leakage of A never exceeds that of B, for any prior π and any gain function g. (We denote this by $A \leq_{\mathcal{G}} B$.)*

Proof. Note first that because $\mathcal{L}_g(\pi, C) = \log {}^{V_g(\pi,C)}/_{V_g(\pi)}$ and $V_g(\pi, C)$ and $V_g(\pi)$ are positive, we have $\mathcal{L}_g(\pi, A) \leq \mathcal{L}_g(\pi, B)$ iff $V_g(\pi, A) \leq V_g(\pi, B)$.

Now

$$V_g(\pi, C) \quad = \quad \sum_{y \in \mathcal{Y}} \max_{w \in \mathcal{W}} \sum_{x \in \mathcal{X}} \pi[x] C_{x,y} g(w, x) \quad ,$$

and as noted in Section 4.C of [1], we can *reify* the choice of w, given y, as a probabilistic channel S from $\mathcal{Y}$ to $\mathcal{W}$ that represents the adversary's *strategy*.[16] Hence we have

$$V_g(\pi, C) \;=\; \max_S \sum_{x,y,w} \pi[x] C_{x,y} S_{y,w}\, g(w, x) \;=\; \max_S \sum_{x,w} \pi[x] (CS)_{x,w}\, g(w, x). \qquad (1)$$

(It might appear that the "max" in equation (1) should actually be "sup," since there are infinitely many possible strategies. But this is not so, because the supremum is in fact realized on any strategy S such that $S_{y,w} > 0$ only if w is a best guess given output y.)

Now notice that in the case where $A = BR$, any optimal strategy S for A is *equivalent to* a strategy for B, namely RS; but of course RS might not be optimal for B — there might be a better strategy S'. This allows us to calculate

$$\begin{array}{lll} & V_g(\pi, A) & \\ = & \max_S \textstyle\sum_{x,w} \pi[x] (AS)_{x,w} g(w, x) & \text{“Eqn. (1)”} \\ = & \max_S \textstyle\sum_{x,w} \pi[x] (BRS)_{x,w} g(w, x) & \text{“$A = BR$”} \\ \leq & \max_{S'} \textstyle\sum_{x,w} \pi[x] (BS')_{x,w} g(w, x) & \text{“S' can be RS”} \\ = & V_g(\pi, B)\,, & \text{“Eqn. (1)”} \end{array}$$

which gives the inequality $V_g(\pi, A) \leq V_g(\pi, B)$ that we seek. □

5.2 Completeness of $(\sqsubseteq_\circ)$

The completeness condition we establish for $(\sqsubseteq_\circ)$ is that if $A \not\sqsubseteq_\circ B$ then there exists a gain function g and a prior π for which A g-leaks strictly more than B does; this depends on a theorem we prove in §6 below. Put informally, this completeness means that if using our order $(\sqsubseteq_\circ)$ we *criticise* a channel A because it does *not* satisfy $A \sqsubseteq_\circ B$, then we can *justify* our criticism by giving a π and g that shows A's inferiority in a more operational setting.

[14] This result first appeared as Theorem 6.2 of [1], though with a slightly different proof.

[15] Proofs for other leakage measures are similar, and indeed since min-entropy leakage is a special case of g-leakage (end §2), that in particular is a trivial corollary.

[16] This reification is reminiscent of Skolemization. Notice that it is reasonable for S to be probabilistic, since there could be more than one w that is optimal for a given y.

Surprisingly, that completeness criterion for ($\sqsubseteq_\circ$) does not hold wrt min-entropy leakage, even though Thm. 1 suggests that it might. This failure is shown by the following example:

$$A = \begin{array}{|c|cc|}\hline x_1 & 2/3 & 1/3 \\ x_2 & 2/3 & 1/3 \\ x_3 & 1/4 & 3/4 \\ \hline\end{array} \qquad B = \begin{array}{|c|ccc|}\hline x_1 & 1/2 & 1/2 & 0 \\ x_2 & 1/2 & 0 & 1/2 \\ x_3 & 0 & 1/2 & 1/2 \\ \hline\end{array}$$

Although it turns out that the min-entropy leakage of A never exceeds that of B on any prior, still $A \not\sqsubseteq_\circ B$.[17]

5.3 Compositionality

A more formal approach to soundness and completeness would be via compositionality, asking *given $A \sqsubseteq_\circ B$, for what contexts C can we be sure that also $C(A) \sqsubseteq_\circ C(B)$?*

In [2] a simple probabilistic programming language with hidden state is treated, with a relation ($\sqsupseteq$) there that specialises to ($\sqsubseteq_\circ$) here when those programs simulate channels. It is shown there that ($\sqsupseteq$) is the (unique) relation with the properties (soundness) that $A \sqsupseteq B$ implies that the min-entropy leakage of $C(A)$ *never* exceeds the min-entropy leakage of $C(B)$ for any context C in that programming language and any prior, and (completeness) that $A \not\sqsupseteq B$ implies that the min-entropy leakage of $C(A)$ *does* strictly exceed the min-entropy leakage of $C(B)$ for some context C and some prior. In this way the legitimacy of ($\sqsupseteq$) for programs, and hence of ($\sqsubseteq_\circ$) for channels, could be argued based on the utility of (the more restricted) min-entropy leakage, and compositionality.

The techniques for proving completeness in [2] led to the proof of Thm. 9 below.

6 The Coriaceous Property and Its Proof

We now present our second major result, the converse to Theorem 8. It says that the strong g-leakage order implies composition refinement, which intuitively means that composition refinement is not *too strong*: that is, whenever $A \not\sqsubseteq_\circ B$, there exists a prior π and a gain function g that causes A to leak more than B. This implication was studied in [1], but not proved in full generality—it was shown only in the case when the columns of B are linearly independent—and the general result was left as the *Coriaceous Conjecture*, which we now resolve.[18]

Theorem 9. *For any channel matrices A and B, if $A \leq_\mathcal{G} B$ then $A \sqsubseteq_\circ B$.*

Proof. We argue the contrapositive, showing that if $A \not\sqsubseteq_\circ B$, then we can construct a gain function g and a prior π such that $V_g(\pi, A) > V_g(\pi, B)$; note that this implies that $\mathcal{L}_g(\pi, A) > \mathcal{L}_g(\pi, B)$ and hence that $A \not\leq_\mathcal{G} B$.

[17] The min-entropy leakage bound can be verified using the linear-programming-based algorithm given in Section 6.F of [1]. To see that $A \not\sqsubseteq_\circ B$, note that because B is invertible we have $A = BR$ implies $R = B^{-1}A$—but this calculation gives an R containing negative entries.

[18] The proof is based on [14], itself extracted from the completeness proof in [2] which was, in turn, a specialisation of McIver's original proof in terms of probabilistic imperative-program fragments and their weakest preconditions [15].

Let A go from $\mathcal{X}$ to $\mathcal{Z}$, and B from $\mathcal{X}$ to $\mathcal{Y}$. If $A \not\sqsubseteq_\circ B$, then there exists no channel matrix R from $\mathcal{Y}$ to $\mathcal{Z}$ such that $A = BR$. If we use the abbreviation $B^\uparrow$ for the matrices $\{BR \mid R \text{ is a channel matrix from } \mathcal{Y} \text{ to } \mathcal{Z}\}$, then our assumption becomes $A \notin B^\uparrow$.

Because matrix A and the matrices in $B^\uparrow$ go from $\mathcal{X}$ to $\mathcal{Z}$, they can be embedded into Euclidean space of dimension $N = |\mathcal{X}|\times|\mathcal{Z}|$ by gluing their columns together in order. Then $B^\uparrow$ becomes a set of points in N-space which, we observe by linearity of matrix multiplication, is both convex and closed. And A is a point in N-space that does not belong to $B^\uparrow$.

By the *Separating Hyperplane Lemma* [16] there is thus a hyperplane in N-space with point A strictly on one side, and all of the set $B^\uparrow$ strictly on the other side. If G is the normal of the hyperplane, also an N-vector thus, we have that $A \cdot G > B' \cdot G$ for all $B' \in B^\uparrow$.[19] Note that we can assume a $(>)$-separation without loss of generality, because we can negate G if necessary. Moreover we can assume without loss of generality that the elements of G are in $[0, 1]$. First, we can eliminate negative elements of G by adding a constant k to each entry; this has the effect of increasing both sides of the inequalities above by exactly $k|\mathcal{X}|$, because with A and each B' derived from "glued" channel matrices, as vectors they all sum to the same value $|\mathcal{X}|$. Second, we can eliminate elements of G that are greater than 1 by scaling, which simply scales both sides of $(<)$ equally.

Now by "ungluing" we can view G, a vector in N-space, as a matrix (though not necessarily a channel matrix) from $\mathcal{X}$ to $\mathcal{Z}$. Thus we can view G as a *gain function* $g : \mathcal{Z}\times\mathcal{X} \to [0, 1]$, using $\mathcal{Z}$ as the set of guesses and defined by $g(z, x) = G_{x,z}$.[20]

It turns out that this g is precisely the gain function that causes A to leak more than B under the uniform prior π_u. For by Eqn. (1) we have

$$\begin{array}{rlcl} & V_g(\pi_u, A) & = & \max_{S_A} \sum_{x,z} \pi_u[x](AS_A)_{x,z}g(z, x) \\ \text{and} & V_g(\pi_u, B) & = & \max_{S_B} \sum_{x,z} \pi_u[x](BS_B)_{x,z}g(z, x) \quad , \end{array}$$

where strategies S_A for A are channel matrices from $\mathcal{Z}$ to $\mathcal{Z}$, and strategies S_B for B are channels matrices from $\mathcal{Y}$ to $\mathcal{Z}$. Note then that the *identity matrix* I is a strategy for A, and that each $BS_B \in B^\uparrow$. Hence, letting S^o_B denote any optimal strategy for B, we have

$$\begin{array}{lll} & V_g(\pi_u, B) & \\ = & \sum_{x,z} \pi_u[x](BS^o_B)_{x,z}g(z, x) & \text{“}S^o_B \text{ is optimal”} \\ = & {}^1\!/_{|\mathcal{X}|} \sum_{x,z}(BS^o_B)_{x,z}G_{x,z} & \text{“}\pi_u \text{ is uniform over } \mathcal{X}\text{”} \\ = & {}^1\!/_{|\mathcal{X}|}\,(BS^o_B) \cdot G & \text{“taking dot-product in vector form”} \\ < & {}^1\!/_{|\mathcal{X}|}\,A \cdot G & \text{“separation; } BS^o_B \in B^\uparrow\text{”} \\ = & \sum_{x,z} \pi_u[x](AI)_{x,z}g(z, x) & \text{“}I \text{ is identity”} \\ \leq & \max_{S_A} \sum_{x,z} \pi_u[x](AS_A)_{x,z}g(z, x) & \text{“}S_A \text{ can be } I\text{”} \\ = & V_g(\pi_u, A) \quad . & \text{“definition } V_g\text{”} \end{array}$$

□

While Theorem 9 shows that composition refinement is no stronger than the strong g-leakage order, one might nonetheless wonder whether the gain function g constructed in the proof (using the Hyperplane Separating Lemma) represents a "practical" leakage

[19] We are using the vector forms here, and $(\cdot)$ is used for their dot-products.

[20] Note that this is the *transpose* of the matrix representation of gain functions used in [1].

threat, in that a "real" adversary would ever care about it. That is, perhaps the strong g-leakage ordering is itself too strong. Three comments seem relevant here. First, it seems generally prudent to make as few assumptions about the adversary as possible. Second, the partial proofs[21] in [1] show that, in the special case when $A \not\sqsubseteq_\circ B$ and the columns of B are linearly independent, there is a quite intuitive gain function g and prior π that causes A to leak more than B; g can then be a *two-block gain function*, which corresponds to the adversary wanting to guess some *property* of the secret. And finally (§5.3), with suitable definition of context it could be possible to reduce ($\sqsubseteq_\circ$) to the strong min-entropy leakage order.

7 The Mathematical Structure of Hyper-distributions

In this section, we give a monadic presentation of composition refinement which, while not necessary for the results in this paper, supports generalisation to richer settings.

7.1 Use of the Giry Monad

In Def. 2 we defined abstract channels as mappings from priors to hyper-distributions. Recall that our (finite) input space is $\mathcal{X}$, and write $\mathbb{D}\mathcal{X}$, with typical element lower-case Greek (e.g. δ, π), for the (discrete) distributions over $\mathcal{X}$; in that case (discrete) *hyper*-distributions have type $\mathbb{D}^2\mathcal{X}$, with typical element upper-case Greek (e.g. Δ), and abstract channels have type $\mathbb{D}\mathcal{X} \to \mathbb{D}^2\mathcal{X}$.[22] We now look at $\mathbb{D}^2\mathcal{X}$ specifically, from a monadic perspective [17].[23]

The functor $\mathbb{G}$ of the Giry monad [19] $(\mathbb{G}, \boldsymbol{\mu}, \boldsymbol{\eta})$ takes a measure space to another space of measures, on the measures of that first space: this is the general technique that allows us to construct distributions $\mathbb{D}()$ "on top of" another set of distributions $\mathbb{D}\mathcal{X}$, as in $\mathbb{D}^2\mathcal{X}$ (and even $\mathbb{D}^3\mathcal{X}$ as in §7.3 below). As part of the monad structure we have a "multiply" natural transformation $\boldsymbol{\mu}$ that averages a distribution of distributions to create a single distribution again. (We see an example of this below.) Here we call it avg for "average." The "unit" natural transformation $\boldsymbol{\eta}$ makes a point distribution on a distribution; but we will not need it here. The functor $\mathbb{G}$ itself, acting on a mapping f e.g. from $\mathcal{X}$ to $\mathcal{Y}$, constructs a "lifted" mapping $\mathbb{G}f$ from $\mathbb{G}\mathcal{X}$ to $\mathbb{G}\mathcal{Y}$, that is in our simple setting from $\mathbb{D}\mathcal{X}$ to $\mathbb{D}\mathcal{Y}$. We call it map here.[24] Finally, we have a function exp that takes the expected value of a function from a measure space to a weighted sum based on a particular measure in that space; we see an example of that immediately below (§7.2).

[21] See the proofs of Lemma 6.4 and Theorems 6.5 and 6.6 of [1].

[22] Since $\mathbb{D}\mathcal{X}$ is uncountable even for finite $\mathcal{X}$, hyper-distributions are at least potentially proper measures: but when derived from matrices, as they are here, they are discrete distributions. The proper-measure case is treated in [11,17] as mentioned in §7.3 below.

[23] We keep this treatment very light: more details are found in [17], where the Kantorovich monad [18] is used in a similar style.

[24] In elementary probability it is called "push forward." Calling it map is by analogy with the use of monads in functional programming, where map "lifts" a function f between elements to a function $\mathsf{map\ f}$ between structures on those elements.

7.2 Applying g-vulnerability to Hyper-distributions Directly

We recall from §2 that a gain function $g: \mathcal{W}\times\mathcal{X} \to [0,1]$ gives rise to two derived functions: the prior vulnerability V_g takes one argument, having type $\mathbb{D}\mathcal{X}\to[0,1]$. The expected vulnerability (again) V_g of the posterior distributions takes two arguments, a prior *and* a channel; but in the mathematical presentation we consider that to be of type $\mathbb{D}^2\mathcal{X}\to[0,1]$, i.e. to have as its *single* argument the hyper-distribution that the prior and channel jointly determine.[25] That is, with this overloading of the name "V_g" it is type-correct to write both $V_g(\delta)$ and $V_g(\Delta)$ for $\delta: \mathbb{D}\mathcal{X}$ and $\Delta: \mathbb{D}^2\mathcal{X}$.

The second form of V_g, applied to a particular hyper-distribution $\Delta: \mathbb{D}^2\mathcal{X}$, is then the expected value $\mathsf{exp}_{V_g}(\Delta)$ over Δ of the first form of V_g as a random variable on $\mathbb{D}\mathcal{X}$.[26]

7.3 Applying Composition Refinement ($\sqsubseteq_\circ$) to Hyper-distributions Directly

We now introduce *bi-hypers* on $\mathcal{X}$, that is hyper-distributions on $\mathbb{D}\mathcal{X}$ (rather than on $\mathcal{X}$ directly), that thus have type $\mathbb{D}^3\mathcal{X}$ with typical element bold upper-case Greek (e.g. $\boldsymbol{\Delta}$). The definition of composition refinement ($\sqsubseteq_\circ$) on hyper-distributions is then as follows:

Definition 5. *Given two hyper-distributions $\Delta_{A,B}: \mathbb{D}^2\mathcal{X}$, we say that $\Delta_A \sqsubseteq_\circ \Delta_B$ just when there is a bi-hyper $\boldsymbol{\Delta}: \mathbb{D}^3\mathcal{X}$ such that*

$$\Delta_A = \mathsf{map}(\mathsf{avg})(\boldsymbol{\Delta}) \quad \textit{and} \quad \mathsf{avg}(\boldsymbol{\Delta}) = \Delta_B\ . \quad [27]$$

The bi-hyper $\boldsymbol{\Delta}$ is thus a witness of the relationship ($\sqsubseteq_\circ$), just as R is a witness in the matrix setting.

This more general, abstract construction of Def. 5 is not necessary for the material (elsewhere) in this paper; but its being expressed purely in monadic terms means it applies without change to proper measures (rather than only discrete distributions). These can arise naturally in a context more general than channels, for example imperative looping programs with hidden state [11], and probabilistic- and demonic nondeterminism together [17]. In this way, the channel model can be seen to fit into this very general mathematical framework, possibly giving access to more general mathematical tools in the analysis of channels.

8 Limits of the Information-Theoretic Perspective

The perspective of abstract channels is information theoretic, concerned only with a channel's mapping from priors to hyper-distributions, and abstracting from details like the names of outputs. These choices are appropriate if we are interested only in the

[25] In [17] the prior vulnerability function is abstracted from any g, presented simply as a "disorder test" that is by definition some continuous, concave function in $\mathbb{D}\mathcal{X}\to[0,1]$. Continuity requires a metric, or a topology, and that is part of what the general Giry- or Kantorovich monad structure supplies. Thus disorder tests are concave (by definition), while g-vulnerabilities are convex (by construction based on g). The latter is a just special case of the former, negated.

[26] That expected value would be written $\int V_g \,\mathrm{d}\Delta$ or $\int_{\delta\in\mathbb{D}\mathcal{X}} V_g(\delta)\,\mathrm{d}\Delta$ in a more mathematical setting.

[27] In the usual notation of the Giry monad that would be $\Delta_A = \mathbb{G}\boldsymbol{\mu}_{\mathcal{X}}\boldsymbol{\Delta}$ and $\boldsymbol{\mu}_{\mathbb{G}\mathcal{X}}\boldsymbol{\Delta} = \Delta_B$.

information that a channel provides to the adversary, and not in the *amount of computation* that might be required in order to exploit that information.

But if we wish to consider computationally-bounded adversaries, then we need to move to a more concrete model, one where outputs come as *strings of bits*. Also, we need to constrain the strategy-based formulation of g-vulnerability that we used in the proof of Theorem 8. For simplicity, let us restrict our attention to min-entropy leakage and (ordinary) vulnerability, whose strategy-based formulation is

$$V(\pi, C) \quad = \quad \max_S \sum_{x,y} \pi[x] C_{x,y} S_{y,x} \quad .$$

In a computational setting, we can no longer allow S to be an arbitrary probabilistic mapping from outputs $\mathcal{Y}$ to guesses $\mathcal{X}$, but instead must require it to be efficiently computable. This in turn requires that we consider *families* of channels with respect to a "security parameter" n, so that we can consider the growth of running time as a function of n. Let us write V^c to denote the computational version of vulnerability.[28]

We can illustrate the effect of this definition by considering two channels whose input is an n-bit prime p, assumed uniformly distributed. Channel A outputs p^2, while channel B outputs pq, where q is a uniformly-distributed $(n+1)$-bit prime. Note that A and B represent the *same* abstract channel, since the reduced matrix of both is the identity matrix. Hence in the non-computational setting we have $V(\pi, A) = V(\pi, B) = 1$.

Turning next to V^c, we find that $V^c(\pi, A) = 1$, since there is an efficient strategy that maps p^2 to p by calculating the square root via binary search. In contrast, $V^c(\pi, B)$ should be smaller, since the existence of an efficient strategy that maps pq to p would contradict the standard assumptions about the difficulty of the factorization problem. Indeed, it would appear that $V^c(\pi, B) \approx V^c(\pi)$, since an efficient probabilistic strategy is believed to have a negligible probability of recovering p from pq.

Here we also have $A \sqsubseteq_\circ B$, which implies by Theorem 8 that $V(\pi, A) \leq V(\pi, B)$. Why does the same inequality not hold for V^c? Recall that the proof of Theorem 8 is based on the fact that if $A = BR$, then any strategy S for A gives rise to an equivalent strategy for B, namely RS. But notice that RS need not be efficiently computable, even if S is. Since here R is a channel that maps pq to p^2, it indeed does not give rise to an efficiently computable strategy for B. In the computational setting, however, we should be able to get a weaker version of Theorem 8 saying that if $A = BR$, where R is *efficiently computable*, then A never out-leaks B.

9 Related Work

Given the multitude of plausible ways to measure the "uncertainty" of a probability distribution and the "amount" of information leakage caused by a channel, there has long been interest in the *robustness* of such measures and the leakage orderings on channels that they give.

[28] There is also a technical issue that arises with *prior vulnerability*. Since now we have a family $\pi^{(n)}$ of priors, parameterized by n, it is not clear that an adversary can efficiently compute an x with maximum probability in $\pi^{(n)}$. In the example that follows, this is not in fact a problem, since there are standard techniques for efficiently generating uniformly-distributed n-bit primes. But in general, we might wish to impose constraints on $\pi^{(n)}$.

Such studies can both establish and refute relationships among measures. For instance, Massey [4] compares Shannon entropy H and guessing entropy G, showing that $G(\pi) > 2^{H(\pi)-2}$, but that there is no interesting upper bound on $G(\pi)$ in terms of $H(\pi)$. Another negative result is given by Pliam [20], who shows the incomparability of Shannon entropy and *marginal guesswork*, which is the minimum number of brute-force guesses required to guess a secret with some specified probability of success. With respect to vulnerability and min-entropy, Santhi and Vardy [21] prove a bound between posterior Shannon entropy and *Bayes risk*, which is the complement of posterior vulnerability; in our notation their bound can equivalently be written as $H(\pi, C) \geq -\log V(\pi, C) = H_\infty(\pi, C)$. Further study of similar bounds is done by Chatzikoklakis, Palamidessi, and Panangaden [22].

Turning to comparisons between channels, we have the results of Yasuoka and Terauchi [7] and Malacaria [8] described in Section 1 that establish the robustness of partition refinement in comparing *deterministic* channels. For *probabilistic channels*, Braun, Chatzikokolakis, and Palamidessi [23] compare the leakage ordering resulting from *multiplicative* and *additive* versions of min-entropy leakage—multiplicative leakage is based on the *ratio* of the posterior- and prior vulnerabilities (as in min-entropy leakage, which is just the logarithm of this ratio), while additive leakage is based on their *difference*. They show that when comparing two channels on a *given* prior, it makes no difference whether multiplicative or additive leakage is used. But when channels are compared with respect to their *capacity* (i.e. maximum leakage over all priors) then multiplicative and additive leakage can produce inconsistent results.

Finally, Sabelfeld and Sands [24] describe a "PER" model of security specifications, based on partitions of the hidden-value space; and there are some similarities between their treatment of partitions and ours: in particular, refining a PER that specifies a program's input could be construed as allowing the program to be less secure; and refining an output PER would require the program to be more secure. Their extension to probability, however, does not seem to lead to the same relation between channels as our does.

10 Conclusion

This paper can be seen as an exploration of the mathematical foundations of quantitative information flow. We have argued that, from the information-theoretic perspective, it is *abstract channels* that are the fundamental objects of study: for when we consider the information-theoretic leakage caused by a channel C, the essential fact is precisely the mapping that C gives from priors to hyper-distributions—and any of the multitude of possible leakage measures can be seen as simply *summarizing* this mapping. Concretely, then, we have seen that classical channel matrices contain structural redundancies which ought to be quotiented away, leading to *reduced matrices*. The utility of the abstract-channel framework is further clarified by our study of *composition refinement*, which is only a pre-order on channel matrices, but which we have proved is a *partial order* on abstract channels. And, by our proof that composition refinement coincides with the *strong g-leakage ordering*, it is a partial order with both structural- and leakage-testing significance—and is therefore a compelling generalization (from deterministic

to probabilistic channels) of *partition refinement* in the Lattice of Information. Finally, we have discussed the limits of the information-theoretic perspective, pointing out that the abstract channels framework is not suitable for addressing computationally-bounded adversaries.

We have shown that channels can be regarded as functions from priors to hyper-distributions and sketched in §7 how they can be formalised using general mathematical machinery; in future work we will investigate this abstraction further in its relation to channels. The characterisation of hypers within the general type of functions would be the first step towards determining which program contexts preserve the order. For example, the Coriaceous result establishes how to show that two channels are not related by ($\sqsubseteq_\circ$) by finding a refuting gain function g; an interesting result would be to determine whether this g can be used to produce the precise conditions under which e.g. min-entropy testing would fail, in the style of program testing "in context" [2]. Another interesting question is whether two programs with an abstract channel denotation can be proved to be in the ($\sqsubseteq_\circ$) relation based on examining the way in which they were constructed. Similar ideas have been discussed in [25] for the specific case of preserving a particular threshold of leakage with respect to a single entropy measurement.

More generally, since particular leakage measures are appropriate for particular applications, we can define a family of weaker pre-orders on abstract channels for a fixed leakage measure m: we say $\mathcal{A} \leq_m \mathcal{B}$ iff the m-leakage of $\mathcal{A}$ never exceeds that of $\mathcal{B}$, for any prior π. What we do not know is whether these are *partial orders* for important choices of m, such as Shannon-, guessing-, or min-entropy leakage. Nor do we know whether they are *strictly weaker* than ($\sqsubseteq_\circ$), though we do know this for $\leq_{min-entropy}$ by the example in §5.2.

Finally, our preliminary investigations suggest that ($\sqsubseteq_\circ$) is not a lattice [26]; future work will reveal other general properties and how to exploit them in channel analysis.

Acknowledgments. Geoffrey Smith and Barbara Espinoza were partially supported by the National Science Foundation under grant CNS-1116318. McIver and Morgan were supported by the Australian Research Council under grant DP120101413. Finally, this paper builds on work [1] done jointly with Mário Alvim, Kostas Chatzikokolakis, and Catuscia Palamidessi, to whom we are deeply appreciative.

References

1. Alvim, M.S., Chatzikokolakis, K., Palamidessi, C., Smith, G.: Measuring information leakage using generalized gain functions. In: Proc. 25th IEEE Computer Security Foundations Symposium (CSF 2012), pp. 265–279 (June 2012)
2. McIver, A., Meinicke, L., Morgan, C.: Compositional closure for Bayes risk in probabilistic noninterference. In: Abramsky, S., Gavoille, C., Kirchner, C., Meyer auf der Heide, F., Spirakis, P.G. (eds.) ICALP 2010. LNCS, vol. 6199, pp. 223–235. Springer, Heidelberg (2010)
3. Shannon, C.E.: A mathematical theory of communication. Bell System Technical Journal 27, 379–423, 623–656 (1948)
4. Massey, J.L.: Guessing and entropy. In: Proc. 1994 IEEE International Symposium on Information Theory, p. 204 (1994)
5. Smith, G.: On the foundations of quantitative information flow. In: de Alfaro, L. (ed.) FOSSACS 2009. LNCS, vol. 5504, pp. 288–302. Springer, Heidelberg (2009)

6. Landauer, J., Redmond, T.: A lattice of information. In: Proc. 6th IEEE Computer Security Foundations Workshop (CSFW 1993), pp. 65–70 (June 1993)
7. Yasuoka, H., Terauchi, T.: Quantitative information flow — verification hardness and possibilities. In: Proc. 23rd IEEE Computer Security Foundations Symposium (CSF 2010), pp. 15–27 (2010)
8. Malacaria, P.: Algebraic foundations for information theoretical, probabilistic and guessability measures of information flow. CoRR abs/1101.3453 (2011)
9. McIver, A., Meinicke, L., Morgan, C.: Compositional closure for Bayes risk in probabilistic noninterference. CoRR abs/1007.1054 (2010) (Draft full version of [2] with appendices)
10. Cover, T.M., Thomas, J.A.: Elements of Information Theory, 2nd edn. John Wiley & Sons, Inc. (2006)
11. McIver, A., Meinicke, L., Morgan, C.: Hidden-Markov program algebra with iteration. At arXiv:1102.0333v1 (2011) (To appear in Mathematical Structures in Computer Science in 2012)
12. Desoer, C.A.: Communication through channels in cascade. PhD thesis, Massachusetts Institute of Technology (1953)
13. Gallager, R.G.: Information Theory and Reliable Communication. John Wiley & Sons, Inc. (1968)
14. McIver, A., Meinicke, L., Morgan, C.: Draft proof of the Coriaceous Conjecture (November 2012), http://www.dagstuhl.de/mat/index.en.phtml?12481
15. McIver, A., Morgan, C.: Abstraction, Refinement and Proof for Probabilistic Systems. Technical Monographs in Computer Science. Springer, New York (2005)
16. Trustrum, K.: Linear Programming. Library of Mathematics. Routledge and Kegan Paul, London (1971)
17. McIver, A., Meinicke, L., Morgan, C.: A Kantorovich-monadic powerdomain for information hiding, with probability and nondeterminism. In: Proc. 27th IEEE Symposium on Logic in Computer Science (LICS 2012), pp. 461–470 (2012)
18. van Breugel, F.: The metric monad for probabilistic nondeterminism (2005), Draft available at http://www.cse.yorku.ca/~franck/research/drafts/monad.pdf
19. Giry, M.: A categorical approach to probability theory. In: Categorical Aspects of Topology and Analysis. Lecture Notes in Mathematics, vol. 915, pp. 68–85. Springer (1981)
20. Pliam, J.O.: On the incomparability of entropy and marginal guesswork in brute-force attacks. In: Roy, B., Okamoto, E. (eds.) INDOCRYPT 2000. LNCS, vol. 1977, pp. 67–79. Springer, Heidelberg (2000)
21. Santhi, N., Vardy, A.: On an improvement over Rényi's equivocation bound. In: 44th Annual Allerton Conference on Communication, Control, and Computing (2006)
22. Chatzikokolakis, K., Palamidessi, C., Panangaden, P.: On the Bayes risk in information-hiding protocols. Journal of Computer Security 16(5), 531–571 (2008)
23. Braun, C., Chatzikokolakis, K., Palamidessi, C.: Quantitative notions of leakage for one-try attacks. In: Proc. 25th Conference on Mathematical Foundations of Programming Semantics (MFPS 2009). ENTCS, vol. 249, pp. 75–91 (2009)
24. Sabelfeld, A., Sands, D.: A PER model of secure information flow. Higher-Order and Symbolic Computation 14(1), 59–91 (2001)
25. Braun, C., Chatzikokolakis, K., Palamidessi, C.: Compositional methods for information-hiding. In: Amadio, R.M. (ed.) FOSSACS 2008. LNCS, vol. 4962, pp. 443–457. Springer, Heidelberg (2008)
26. McIver, A., Morgan, C., Meinicke, L., Smith, G., Espinoza, B.: Abstract channels, gain functions and the information order. In: FCS 2013 Workshop on Foundations of Computer Security (2013), http://prosecco.gforge.inria.fr/personal/bblanche/fcs13/fcs13proceedings.pdf