A new measure for dynamic leakage based on quantitative information flow

Luigi D. C. Soares
UFMG & Macquarie University
Belo Horizonte, Brazil / Sydney, Australia

Mário S. Alvim
UFMG
Belo Horizonte, Brazil

Natasha Fernandes
Macquarie University
Sydney, Australia

***Abstract*—Quantitative information flow (QIF) is concerned with assessing the leakage of information in computational systems. In QIF there are two main perspectives for the quantification of leakage. On one hand, the *static perspective* considers all possible runs of the system in the computation of information flow, and is usually employed when preemptively deciding whether or not to run the system. On the other hand, the *dynamic perspective* considers only a specific, concrete run of the system that has been realised, while ignoring all other runs. The dynamic perspective is relevant for, e.g., system monitors and trackers, especially when deciding whether to continue or to abort a particular run based on how much leakage has occurred up to a certain point. Although the static perspective of leakage is well-developed in the literature, the dynamic perspective still lacks the same level of theoretical maturity. In this paper we take steps towards bridging this gap with the following key contributions: (i) we provide a novel definition of dynamic leakage that decouples the adversary's *belief* about the secret value from a *baseline* distribution on secrets against which the success of the attack is measured; (ii) we demonstrate that our formalisation satisfies relevant information-theoretic axioms, including non-interference and relaxed versions of monotonicity and the data-processing inequality (DPI); (iii) we identify under what kind of analysis strong versions of the axioms of monotonicity and the DPI might not hold, and explain the implications of this (perhaps counter-intuitive) outcome; (iv) we show that our definition of dynamic leakage is compatible with the well-established static perspective; and (v) we exemplify the use of our definition on the formalisation of attacks against privacy-preserving data releases.**

***Index Terms*—quantitative information flow, dynamic leakage**

I. Introduction

The field of *quantitative information flow (QIF)* is concerned with assessing the amount of information that flows through a system. This assessment is relevant both in cases in which the flow of information through the system is desirable (e.g., in a census dataset which produces output statistics reflecting trends in the population data fed as input) and (more commonly) in cases where the flow of information is undesirable (e.g., in electronic voting protocols in which the final tally should not reveal any undue information about the ballot of each voter). QIF has been successfully applied to a wide variety of scenarios, including privacy and security [1, 2, 3, 4, 5, 6, 7], machine learning [8, 9], robotics [10], recommendation systems [11], fairness [12], among others.

In the QIF framework a system is modelled as an information-theoretic channel taking in some secret input and producing some observable output. Attackers are Bayesian, modelled using a prior knowledge and a gain function describing their goals and capabilities. By observing the channel's output, the adversary can update their prior distribution on secret values to a collection of posterior, revised distributions, each occurring with its own probability. The information leakage is defined as the difference between the "vulnerability" of (i.e., the "information" about) the secret, from the point of view of the attacker, before and after passing through the channel (prior and posterior vulnerabilities, respectively). These ideas are captured in the *g-leakage* framework [13].

It is known [14, 15] that the vulnerability measures in the g-leakage framework satisfy some fundamental information-theoretic axioms. Among these, three are of particular interest to this paper, as they deal with the conservation of information in a system. The first, *monotonicity*, says that, by observing the behaviour of the system, the adversary can never end up with less information about the system's input than they had before.[1] The second, known as the *data-processing inequality (DPI)*, says that the post-processing of a system's output cannot increase the information the adversary has about the system's input.[2] Lastly, *non-interference* says that a system that never produces any useful output cannot leak information.

These axioms are satisfied under the *static perspective* of the g-leakage framework, which considers *all possible outputs* of the system, under either *expectation* (by averaging the vulnerability over all posteriors) or *max-case* (by taking the worst-case vulnerability over all posteriors) [14, 15]. Static leakage is useful for analysing a system before it is executed, e.g., when preemptively deciding whether or not to run it.

However, there are situations in which it is desirable to measure leakage with respect to a single run of the system, also known as the *dynamic* perspective. In this case only the posterior corresponding to the realised output is considered, and all posteriors corresponding to non-realised outputs are ignored. The dynamic perspective is particularly useful in leakage trackers or monitors, which follow in real time the information leaked by a system and may allow or abort further execution based on the leakage that has occurred so far.

[1] This is akin to Shannon's *"information can't hurt"* property for entropies, which means conditioning cannot increase entropy: $H(X \mid Y) \leq H(X)$ [16].

[2] This is a generalisation of Shannon's property of DPI that states that if $X \rightarrow Y \rightarrow Z$ form a Markov chain, then $H(X \mid Y) \leq H(X \mid Z)$ [16].

Although of significant interest, the literature lacks a sound definition for dynamic leakage that enjoys the properties of static leakage identified earlier. More precisely, the "traditional" definition of dynamic leakage [15, 17], motivated by the static definition of leakage, simply compares the posterior vulnerability of a secret (computed from a specific observation) with the prior vulnerability. However, this definition may lead to unexpected and incorrect conclusions, as we cover in Section II. Indeed, the QIF community is aware of the need for a more principled definition of dynamic leakage [17].

Contributions. In this paper we propose a definition of dynamic information leakage that fills this gap in the QIF literature. Our work makes the following key contributions:

1) We provide a novel definition of dynamic leakage that is parameterised by two states of knowledge (probability distributions): one modelling the adversary's *belief* about the secret value — which is used to guide their actions — and another one modelling the *baseline* distribution on secrets — against which the success of the adversary's actions are measured. Such a distinction between states of knowledge has already been made by Clarkson et al. [18], who defined information leakage in terms of the Kullback-Leibler (KL) divergence. We extend their work to more general adversarial models, via the g-leakage framework [19], and generalise the notion of "baseline" from "strict truth" to a state of knowledge that is the "closest to the truth" available.
2) We show that the proposed formalisation satisfies the fundamental axiom of non-interference, and satisfies relaxed versions (which we call "single-step") of the axioms of monotonicity and the data-processing inequality. We also show that additional post-processing steps might have the opposite effect to what is expected, increasing vulnerability (dually, decreasing uncertainty). This (perhaps) counter-intuitive result is also not always captured using the traditional definition of dynamic leakage.
3) We identify under what kind of analysis strong versions (which we call "multi-step") of the axioms of monotonicity and the data-processing inequality might not hold, noting that results obtained under such scenarios do not invalidate — and are in fact not comparable to — the guarantees of single-step monotonicity and DPI. We also explain why, from the analyst's perspective, a break of monotonicity (i.e., a negative amount of leakage) naturally corresponds to privacy improvement.
4) We show that the proposed definition of dynamic measures is consistent with the static perspective. That is, the expectation of dynamic vulnerability (or uncertainty) corresponds to the traditional definition of expected vulnerability (uncertainty), and similar for the max-case.
5) We formalise attacks against privacy-preserving data-release systems using QIF and study the application of the new definition of dynamic vulnerability to evaluate the effect of adding a privacy mechanism to the system.

Outline of the paper. In Section II we explore the failures of monotonicity and DPI in the traditional definition of dynamic leakage, motivating our novel definition. In Section III we provide the current definitions of leakage measures using the QIF framework, both in the static and dynamic perspectives. In Section IV we propose a new definition of dynamic leakage, show the necessary steps to adopt it and its limitations, demonstrate that it satisfies non-interference, and show that it partially satisfies monotonicity and the DPI. We also show that this formalisation is consistent with the static perspective. In Section V we analyse a real-world application of privacy-preserving data releases, in which we compare two systems satisfying the DPI and show that the traditional definition of dynamic leakage is not adequate, while our novel formalisation is. In Section VI we discuss the difference between single- and multi-step analysis, and the implications of additional dynamic steps, which might break the axioms of monotonicity and the DPI. In Section VII we discuss related work. Finally, in Section VIII we conclude and discuss future work. Full proofs of lemmas and theorems are provided in Appendix A.

II. Motivating Examples

In this section we exemplify how the traditional definition of dynamic leakage might lead to misleading conclusions. For that, consider a Bayesian adversary with access to the system of interest, including implementation details, who starts with some prior knowledge about a secret value fed as input to the system and, after observing the system's behaviour, updates their knowledge using Bayes' rule. The amount of information leaked by the system is customarily defined as the increase in the adversary's information about the secret, captured as the change in the vulnerability of the secret — that is, the amount of useful information to perform an attack — before and after the system's execution (or, dually, the decrease in the adversary's uncertainty about the secret).

A. The problem of monotonicity: The traditional definition of dynamic leakage can produce incorrect negative values

Intuitively, we expect vulnerability/uncertainty measures to be *monotonic*, meaning that observing an output should never decrease the attacker's information, and so leakage should always be non-negative. It has already been shown that monotonicity is indeed always preserved in the static perspective, for all measures satisfying a few fundamental information-theoretic axioms [15, Thm 5.8]. However, as shown by Bielova [17], it is possible to obtain negative leakage when considering a traditional definition of dynamic leakage that simply subtracts the posterior uncertainty from the prior uncertainty. To illustrate, we review Bielova's Example 1:

Example 1a. Consider a (deterministic) program that takes a 2-bit secret $X \in \{00, 10, 11\}$ and produces a value $Y \in \{a, b\}$:

$$\text{if } X = 00 \text{ then } Y = a \text{ else } Y = b$$

Now, suppose that the adversary's prior knowledge about the secret is modelled by the distribution $p_X = \{{}^7\!/_8, {}^1\!/_{16}, {}^1\!/_{16}\}$. (We write $p_X = \{p_{x_i}\}_{i=0}^{n-1}$, matching the indices in $\{x_i\}_{i=0}^{n-1}$.) Then, let us say that the program is executed and produces an

output b that rules out secret value 00 with certainty, leading the adversary to a posterior knowledge of $p_{X|b} = \{0, 1/2, 1/2\}$.

A popular measure of uncertainty is the Shannon entropy $H(p)$ of a probability distribution $p : \mathbb{D}\mathcal{X}$, defined as [20]

$$H(p) \stackrel{\text{def}}{=} -\sum_{x \in \mathcal{X}} p_x \log_2 p_x \tag{1}$$

Recall that we expect a *decrease* of the adversary's uncertainty once an output is observed. Hence, in order to have positive values of leakage representing some flow of information, we would typically define the (dynamic) Shannon leakage as

$$H(p_X) - H(p_{X|y}) \tag{2}$$

Therefore, going back to our example, the dynamic leakage caused by observing output b is

$$H(p_X) - H(p_{X|b}) \approx 0.67 - 1 = -0.33$$

implying that the attacker's uncertainty has *increased* after observing that particular output, even though the adversary has effectively learned the first bit of the secret. □

Example 1a can be captured by the *Belief Tracking* framework of Clarkson et al. [18]. Although for deterministic systems that approach does not depend on the concrete value of the secret [17, Thm 1], in general (for probabilistic systems) the framework requires a concrete secret value, and it measures uncertainty with the KL divergence. Our approach allows for more general adversarial models via the g-leakage framework.

B. The problem of DPI: The traditional definition of dynamic leakage is not well-suited for comparing systems

It is often the case that we want to compare two systems C and D in terms of their information leakage. This is especially useful when C and D are related — for example, when D is exactly the program C with an additional code snippet, which can be modelled as a post-processing step. One might include a post-processing step to increase security/privacy; this intuition relies on the *data-processing inequality* (DPI), which says that post-processing cannot *increase* information leakage, and therefore D must be *at least as safe as* C. In QIF we call D a *refinement* of C, written $\mathrm{C} \sqsubseteq \mathrm{D}$, whenever D is a post-processing of C.[3] It has been shown that the static perspective of leakage preserves the DPI, and thus refinement.

One might wonder whether this notion of refinement, i.e., of being able to safely replace a system C with D if system D can be proven to be a post-processing of C, also holds under dynamic scenarios. Unfortunately, by employing current definitions of dynamic information leakage, this fails in many ways. To illustrate that, Example 2a below shows a case in which the results suggest that such a replacement could be performed, but in practice we can verify that it is not necessarily safe. A second example will be detailed in Section V.

Example 2a. Consider a company that collects health data and provides public access to it via queries, such as if there is a case of a particular disease in the database. Due to the sensitive nature of the data, to ensure individuals' privacy, the data engineering team develops the following program P which produces as output a value $y \in \{\text{NO}, \text{YES}\}$, where $(a \ {}_p\!\oplus\ b)$ returns a with probability p and b with $1 - p$:

$$\text{if answer = NO then } (\text{NO } {}_{2/3}\!\oplus\ \text{YES}) \text{ else } (\text{NO } {}_{1/3}\!\oplus\ \text{YES})$$

That is, this program reports the true answer with ⅔ probability and the wrong answer with ⅓ probability.

To test P, the data engineering team runs a query from the perspective of an attacker $\mathscr{A}^{\mathrm{P}}$ and measures the chance of correctly inferring the original answer. Assume that the adversary starts with no prior knowledge, modelled as a uniform distribution $p_X = \{1/2, 1/2\}$. Then the value YES is observed by $\mathscr{A}^{\mathrm{P}}$, who uses Bayes' rule to obtain the posterior $p^{\mathscr{A}^{\mathrm{P}}}_{X|\text{YES}} = \{1/3, 2/3\}$. Thus, $\mathscr{A}^{\mathrm{P}}$'s best action is to guess YES.

To measure the secret's vulnerability with respect to $\mathscr{A}^{\mathrm{P}}$'s action, we use the Bayes vulnerability, which chooses the secret with the highest probability, and corresponds to the adversary's probability of guessing the secret in one try:

$$V_1(p) \stackrel{\text{def}}{=} \max_{x \in \mathcal{X}} p_x \tag{3}$$

Hence, the (dynamic) vulnerability is $V_1(p^{\mathscr{A}^{\mathrm{P}}}_{X|\text{YES}}) = 2/3$.

Now, let us say that the engineering team considers this vulnerability to be too large, and decides to add a second sanitisation step S to the pipeline, implemented as follows:

$$\text{if (perturbed) answer = NO then (NO) else } (\text{NO } {}_{1/2}\!\oplus\ \text{YES})$$

That is, S always passes on input NO accurately, whereas input YES has 1/2 probability of being flipped.

The new post-processing mechanism S is then applied to the output produced by P. In our example, let us say that the output YES produced by P is fed as input to S, which produces an output NO, which is observed by a second adversary $\mathscr{A}^{\mathrm{S}}$ (see Figure 1 for an illustration). We can compute $\mathscr{A}^{\mathrm{S}}$'s posterior knowledge via Bayes' rule as $p^{\mathscr{A}^{\mathrm{S}}}_{X|\text{NO}} = \{5/9, 4/9\}$.

Therefore, $\mathscr{A}^{\mathrm{S}}$'s best action is to guess NO, with a vulnerability of $V_1(p^{\mathscr{A}^{\mathrm{S}}}_{X|\text{NO}}) = 5/9$, which is *smaller* than $\mathscr{A}^{\mathrm{P}}$'s chance of correctly guessing the secret. So, by employing the traditional definition of dynamic vulnerability, it seems that the engineering team has achieved their goal: to come up with a system (the composition P ; S) that is *always* safer than P.

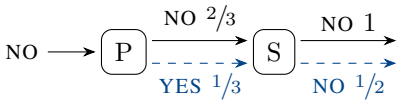


Fig. 1: Pipeline P ; S for Example 2a, considering as input the answer NO and as final output the answer NO. The outgoing black arrows correspond to the case where the mechanism preserved the answer, whereas the outgoing (blue) dashed arrows indicate that the answer was flipped. The probability of an output NO when its input is NO is $2/3 \cdot 1 + 1/3 \cdot 1/2 = 5/6$.

[3] The definition is a little more nuanced than this, but these notions have been shown to be equivalent.

However, a more careful analysis shows that this conclusion can be misleading. Suppose that the actual input to P was NO. This is the real secret, which is known to the data engineering team. Notice that $\mathscr{A}^{\mathrm{P}}$'s action of guessing YES would not match the secret value in this scenario, whereas $\mathscr{A}^{\mathrm{S}}$'s action of guessing NO would. Thus, *in reality* $\mathscr{A}^{\mathrm{S}}$ would guess correctly while $\mathscr{A}^{\mathrm{P}}$ guesses incorrectly, even though we computed that $\mathscr{A}^{\mathrm{S}}$'s chance of success was smaller than $\mathscr{A}^{\mathrm{P}}$'s chance. □

The example above shows that the traditional definition of information leakage can lead to erroneous conclusions under dynamic scenarios. Our goal is then to provide a new definition that captures dynamic scenarios correctly, and also to identify when the data-processing inequality is satisfied.

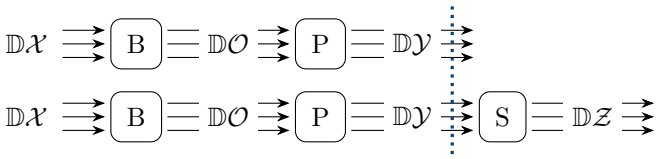
C. A brief explanation on why the traditional definition of dynamic leakage leads to unexpected results

Our key insight is that, in the traditional definition of leakage, the adversary's knowledge about the secret plays two distinct roles at the same time: it represents the adversary's *belief* about the secret value — which is used to guide their actions — but it is also the *baseline* distribution on secrets — against which the success of the adversary's actions are measured. This is a problem in the dynamic perspective because, once a concrete execution of the system is observed, the prior distribution on secrets is updated to a posterior distribution that better captures what values the secret could have taken *in that particular run*. Therefore, although the adversary's prior distribution remains a good reflection of their *belief* about the secret values before the system was run, it is not necessarily an accurate reflection of the *baseline* against which their success should be measured *in that particular run*.

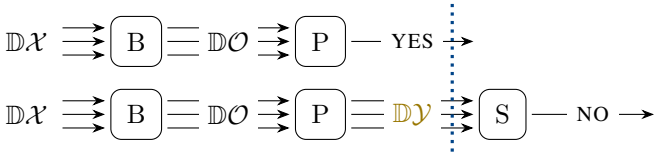
As such, we argue that we should divide the states of knowledge between baseline and belief. As we shall formalise in this paper, the absence of such a distinction did not pose a problem under the static perspective because the expectation of all posteriors — which act as both the adversary's posterior beliefs and also the baselines — is indeed always equal to the prior (belief) and the effect is, therefore, not detectable.

To explore in more detail why the traditional definition of dynamic leakage is not always suitable for comparing systems, let us contrast it to the static perspective. Figures 2a and 2b illustrate what the semantics of, respectively, the static and (traditional) dynamic analysis would be with respect to the adversaries from Example 2a. Recall that adversary $\mathscr{A}^{\mathrm{P}}$ observes the output of the system after the first post-processing step P (represented in the upper diagram in each figure), while adversary $\mathscr{A}^{\mathrm{S}}$ observes the output of the system after the second post-processing step S (lower diagram). The first component of the system in both diagrams is B, taking as input a secret value from a set $\mathcal{X}$ and producing as output a value from a set $\mathcal{O}$, which could, in principle, have been observed by the data engineering team. In this case the component B is simply the program that reports the original, unaltered answer.

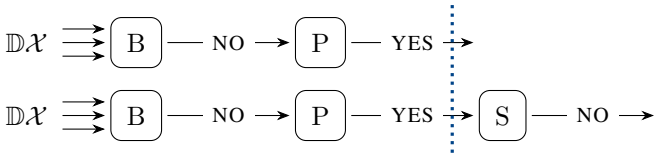
Notice that, in the static perspective (Figure 2a), the section from $\mathbb{D}\mathcal{X}$ up to $\mathbb{D}\mathcal{Y}$ in the lower diagram (adversary $\mathscr{A}^{\mathrm{S}}$)



(a) Static perspective.



(b) Traditional dynamic perspective, when the output observed by adversary $\mathscr{A}^{\mathrm{P}}$ is YES and the output observed by $\mathscr{A}^{\mathrm{S}}$ is NO.



(c) (Multi-step) dynamic perspective, from the point of view of the data engineering team, when the real, secret value is NO.

Fig. 2: The semantics of analyses related to adversaries $\mathscr{A}^{\mathrm{P}}$ (upper diagram in each figure) and $\mathscr{A}^{\mathrm{S}}$ (lower diagram) from Example 2a. In each figure, the vertical dotted line delimits the section of the pipeline that is common to both adversaries. Multiple arrows indicate a "static" input (respectively, output), meaning that the system takes as input (produces as output) a probability distribution over all possible values. A single arrow indicates a concrete, single execution of the system.

exactly matches the upper diagram (adversary $\mathscr{A}^{\mathrm{P}}$). That is, the semantics of both analyses up to the output of system P — the section common to both analyses — takes into account every possible execution path. This indicates that the results of the two analyses in the static perspective are comparable.

In contrast, in the (traditional) dynamic perspective (Figure 2b), the section from $\mathbb{D}\mathcal{X}$ up to $\mathbb{D}\mathcal{Y}$ in the lower diagram ($\mathscr{A}^{\mathrm{S}}$) does *not* match with the upper diagram ($\mathscr{A}^{\mathrm{P}}$). That is, while the analysis of adversary $\mathscr{A}^{\mathrm{P}}$ (upper diagram) takes into account a single output YES from system P, the analysis of $\mathscr{A}^{\mathrm{S}}$ (lower diagram) considers both outputs YES and NO from P. Therefore, we argue that these results are *not* comparable.

The intuition in this case is that the traditional definition of dynamic vulnerability is parameterised by only one explicit state of knowledge: the adversary's belief after observing the final output. However, implicitly, this definition evaluates the effect of the actions taken by the adversary, derived from their belief, on every possible baseline — i.e., every possible state of knowledge that could have been obtained in any intermediate step prior to the output observed by the adversary — as an expectation. The problem with this is then that the set of states of knowledge considered implicitly in the analysis of $\mathscr{A}^{\mathrm{S}}$ may include states of knowledge that were not

considered in the analysis of $\mathscr{A}^{\mathrm{P}}$. Indeed, this is what happens in Example 2a, when considering the traditional definition of dynamic vulnerability: the result obtained for adversary $\mathscr{A}^{\mathrm{S}}$ considers the intermediate state of knowledge obtained given output NO from P as one of the possible baselines, whereas this state of knowledge is not considered in the result of $\mathscr{A}^{\mathrm{P}}$.

The reason why no issues arise in the static perspective (Figure 2a) is that the semantics of the static perspective not only takes into account (implicitly) every possible baseline, it also considers (explicitly) every possible belief that the adversary could construct. Consequently, the set of states of knowledge considered in the analysis without component S (upper diagram) is exactly the same used in the analysis that includes S (lower diagram), up to the output of component P.

A similar argument follows for Example 1a, in which we obtained negative leakage. In that case we would model P as the deterministic program mapping the 2-bit secret to either a or b, and represent the adversary's prior knowledge as a post-processing step $\mathbb{1}$ that maps every output from P to the same value $\perp$ that does not reveal anything. Then, there is only one posterior knowledge obtained from $\mathbb{1}$, which is exactly the prior knowledge. By doing this, we can see that the set of states of knowledge considered in the analysis a priori includes states of knowledge not considered in the analysis a posteriori.

Finally, we concluded Example 2a by noting that there is at least one dynamic path NO $\xrightarrow{\mathrm{P}}$ YES $\xrightarrow{\mathrm{S}}$ NO that shows that it is not necessarily safe to compose systems P;S. This comparison is represented by the diagrams in Figure 2c. Notice that this time the prefix of the lower diagram ($\mathscr{A}^{\mathrm{S}}$) matches with the upper diagram ($\mathscr{A}^{\mathrm{P}}$), and so the results can be compared.

III. Preliminaries

In this paper we employ the mathematical framework of *quantitative information flow* (QIF) [15] to model (probabilistic) systems and characterise Bayesian adversarial attacks against them. In the rest of this section we formalise systems, adversaries with different goals and capabilities, and vulnerability/uncertainty measures under both the static and the dynamic perspectives, using the framework of QIF.

A. Modelling systems and knowledge

In QIF a system is modelled as an *information-theoretic channel* $\mathrm{C} : \mathcal{X} \to \mathbb{D}\mathcal{Y}$ that maps a (secret) input $x \in \mathcal{X}$ to a (public) output $y \in \mathcal{Y}$ according to some distribution in $\mathbb{D}\mathcal{Y}$. (Given a set $\mathcal{Y}$, we denote by $\mathbb{D}\mathcal{Y}$ the set of all probability distributions over $\mathcal{Y}$.) We usually assume that $\mathcal{X}$ and $\mathcal{Y}$ are discrete, in which case we write C as a stochastic matrix with rows labeled by the possible inputs and columns labeled by the possible outputs, so that $\mathrm{C}_{x,y}$ is the probability of channel C outputing value $y \in \mathcal{Y}$ when its input is $x \in \mathcal{X}$.

By combining a *prior* knowledge $\pi : \mathbb{D}\mathcal{X}$ with the channel C that models the system of interest, one can compute a joint distribution $(\pi \triangleright \mathrm{C})_{X,Y} : \mathbb{D}(\mathcal{X} \times \mathcal{Y})$ as

$$(\pi \triangleright \mathrm{C})_{x,y} \stackrel{\text{def}}{=} \pi_x \mathrm{C}_{x,y} \quad \forall x \in \mathcal{X} \text{ and } y \in \mathcal{Y} \tag{4}$$

Then, from the joint distribution, one can derive a marginal distribution $(\pi \triangleright \mathrm{C})_Y : \mathbb{D}\mathcal{Y}$ on the possible observations $y \in \mathcal{Y}$, which we call the *outer* distribution, as

$$(\pi \triangleright \mathrm{C})_y \stackrel{\text{def}}{=} \sum_{x \in \mathcal{X}} (\pi \triangleright \mathrm{C})_{x,y} \quad \forall y \in \mathcal{Y} \tag{5}$$

From the joint and outer distributions, one can construct posterior distributions $(\pi \triangleright \mathrm{C})_{X|y} : \mathbb{D}\mathcal{X}$ on the values $\mathcal{X}$ given each $y \in \mathcal{Y}$, which we also call the *inner* distributions, as

$$(\pi \triangleright \mathrm{C})_{x|y} \stackrel{\text{def}}{=} \frac{(\pi \triangleright \mathrm{C})_{x,y}}{(\pi \triangleright \mathrm{C})_y} \quad \forall x \in \mathcal{X} \tag{6}$$

Finally, when considering a static perspective, we write $[\pi \triangleright \mathrm{C}]$ to denote a *hyper-distribution*, a distribution on distributions on $\mathcal{X}$, (customarily) abstracting away the actual output labels.

Channels can be composed in many ways. Relevant to this paper is the sequential composition of two channels C and D, termed *cascade* and written $\mathrm{C} \,;\, \mathrm{D}$. The cascade of a channel C followed by a channel D is defined as follows:

Definition 1 (Cascade [15, Definition 4.18])**.** Given two channel (matrices) $\mathrm{C} : \mathcal{X} \to \mathbb{D}\mathcal{Y}$ and $\mathrm{D} : \mathcal{Y} \to \mathbb{D}\mathcal{Z}$, the cascade $\mathrm{C} \,;\, \mathrm{D}$ of channels C and D is the channel matrix of type $\mathcal{X} \to \mathbb{D}\mathcal{Z}$, obtained by ordinary matrix multiplication.

B. Modelling adversarial goals

Adversaries can have different goals: they may be interested in guessing the exact secret, guessing the secret within a certain number of tries, they may deem some secret values more valuable than others, etc. Each goal can be captured in the g-leakage framework [19] as a suitable gain function $g : \mathcal{W} \times \mathcal{X} \to \mathbb{R}$, which measures the benefit to the adversary when they take an *action* $w \in \mathcal{W}$ and the secret value is $x \in \mathcal{X}$. For instance, an adversary who wants to guess the secret in one try is modelled by the identity gain function

$$\mathbf{1}(w, x) \stackrel{\text{def}}{=} 1 \text{ if } w = x \text{ else } 0 \tag{7}$$

As we shall see (Section III-C), gain functions are employed in the definition of vulnerability measures. Dually, entropy measures can be defined in terms of loss functions of type $\ell : \mathcal{W} \times \mathcal{X} \to \mathbb{R}_{\geq 0} \cup \{\infty\}$, modelling the attacker's loss upon taking a particular action. For instance, Shannon entropy corresponds to an adversary who is trying to guess the "true" distribution on $\mathcal{X}$ (that is, each action is a distribution). Therefore, it can be modelled by the following loss function:

$$\ell_H(w, x) \stackrel{\text{def}}{=} -\log_2 w_x \tag{8}$$

defining $-\log_2 0 = \infty$,[4] noting that ℓ_H corresponds to Shannon's definition of information content, measured in bits.

Remark 1. We note that, albeit in [15] loss functions are defined disallowing ∞, we find it more useful to include ∞, to permit results on Shannon entropy. This is not a problem, as we still require uncertainty measures to be bounded.

[4] For more examples, see [15, §3.2].

C. Current definitions of vulnerability, uncertainty, and information leakage under static and dynamic perspectives

The threat to a secret caused by an adversary is given by a *g-vulnerability measure* $V_g(p)$, defined as the expected gain to the adversary by taking an optimal action:

$$V_g(p) \stackrel{\text{def}}{=} \max_{w \in \mathcal{W}} \sum_{x \in \mathcal{X}} p_x \, g(w, x) \tag{9}$$

replacing $\max$ with $\sup$ if $\mathcal{W}$ is infinite. Dually, the adversary's uncertainty about the secret is given by an *ℓ-uncertainty measure* $U_\ell(p)$, parameterised by a loss function ℓ:

$$U_\ell(p) \stackrel{\text{def}}{=} \min_{w \in \mathcal{W}} \sum_{x \in \mathcal{X}} p_x \, \ell(w, x) \tag{10}$$

replacing $\min$ with $\inf$ when $\mathcal{W}$ is infinite. For example, Shannon entropy can be obtained from (8) as

$$U_{\ell_H}(p) = \inf_{w \in \mathcal{W}} \sum_{x \in \mathcal{X}} p_x \, (-\log_2 w_x) \tag{11}$$

which is minimised exactly when $w = p$, and equals $H(p)$.

We then define the amount of information available to the adversary before and after the system is run, and the amount of information leaked from the system's execution as follows:

Definition 2 (Prior measurements)**.** Given a prior knowledge modelled by a probability distribution $\pi : \mathbb{D}\mathcal{X}$, the *prior* g-vulnerability is $V_g(\pi)$. Dually, the *prior* ℓ-uncertainty is $U_\ell(\pi)$.

Definition 3 (Traditional dynamic posterior measurements)**.** The *dynamic posterior* g-vulnerability with respect to the adversary's posterior knowledge $(\pi \triangleright \mathrm{C})_{X|y}$, constructed by observing an execution of the system C, is $V_g((\pi \triangleright \mathrm{C})_{X|y})$. Dually, the *dynamic posterior* ℓ-uncertainty is $U_\ell((\pi \triangleright \mathrm{C})_{X|y})$.

Definition 4 (Traditional dynamic leakage)**.** Let π be the adversary's prior knowledge. The amount of information that leaks from the execution of a system $\mathrm{C} : \mathcal{X} \to \mathbb{D}\mathcal{Y}$ upon yielding an output $y \in \mathcal{Y}$ is defined (additively) in terms of prior g-vulnerability and posterior g-vulnerability as

$$\mathcal{L}_g((\pi \triangleright \mathrm{C})_{X|y}) \,\|\, \pi) \stackrel{\text{def}}{=} V_g((\pi \triangleright \mathrm{C})_{X|y}) - V_g(\pi)$$

When considering the attacker's ℓ-uncertainty about the secret, we flip the roles of the prior and posterior measures. That is,

$$\mathcal{L}_\ell((\pi \triangleright \mathrm{C})_{X|y}) \,\|\, \pi) \stackrel{\text{def}}{=} U_\ell(\pi) - U_\ell((\pi \triangleright \mathrm{C})_{X|y})$$

Now, under a static perspective, one might be interested in either the expected or worst case scenarios. This is defined (overloading V_g and U_ℓ) as follows, taking into account all possible outputs that could be produced from the system:

Definition 5 (Static posterior measurements)**.** Let $\pi : \mathbb{D}\mathcal{X}$ be the adversary's prior knowledge and $\mathrm{C} : \mathcal{X} \to \mathbb{D}\mathcal{Y}$ the channel modelling a system. The static posterior g-vulnerability is defined in the expected- and max-case as, respectively,

$$V_g[\pi \triangleright \mathrm{C}] \stackrel{\text{def}}{=} \sum_{y \in \mathcal{Y}} (\pi \triangleright \mathrm{C})_y \, V_g((\pi \triangleright \mathrm{C})_{X|y})$$

$$V_g^{\max}[\pi \triangleright \mathrm{C}] \stackrel{\text{def}}{=} \max_{y \in \mathcal{Y}} V_g((\pi \triangleright \mathrm{C})_{X|y})$$

Dually, the static posterior ℓ-uncertainty is defined in the expected- and min-case as, respectively,

$$U_\ell[\pi \triangleright \mathrm{C}] \stackrel{\text{def}}{=} \sum_{y \in \mathcal{Y}} (\pi \triangleright \mathrm{C})_y \, U_\ell((\pi \triangleright \mathrm{C})_{X|y})$$

$$U_\ell^{\min}[\pi \triangleright \mathrm{C}] \stackrel{\text{def}}{=} \min_{y \in \mathcal{Y}} U_\ell((\pi \triangleright \mathrm{C})_{X|y})$$

Definition 6 (Static leakage)**.** The static g-leakage is defined in the expected- and max-case as, respectively,

$$\mathcal{L}_g(\pi, \mathrm{C}) \stackrel{\text{def}}{=} V_g[\pi \triangleright \mathrm{C}] - V_g(\pi)$$

$$\mathcal{L}_g^{\max}(\pi, \mathrm{C}) \stackrel{\text{def}}{=} V_g^{\max}[\pi \triangleright \mathrm{C}] - V_g(\pi)$$

When considering the attacker's ℓ-uncertainty about the secret, we flip the roles of the prior and posterior measures. That is,

$$\mathcal{L}_\ell(\pi, \mathrm{C}) \stackrel{\text{def}}{=} U_\ell(\pi) - U_\ell[\pi \triangleright \mathrm{C}]$$

$$\mathcal{L}_\ell^{\min}(\pi, \mathrm{C}) \stackrel{\text{def}}{=} U_\ell(\pi) - U_\ell^{\min}[\pi \triangleright \mathrm{C}]$$

Finally, we recall the notion of refinement, which tells us when one channel is at least as safe (i.e., leaks no more) than another, and coincides with the cascading of the original channel and a post-processing channel (Definition 1):

Definition 7 (Refinement)**.** Given channels $\mathrm{C} : \mathcal{X} \to \mathbb{D}\mathcal{Y}$ and $\mathrm{D} : \mathcal{X} \to \mathbb{D}\mathcal{Z}$, we say that C is refined by D, written $\mathrm{C} \sqsubseteq \mathrm{D}$, whenever $V_g[\pi \triangleright \mathrm{C}] \geq V_g[\pi \triangleright \mathrm{D}]$ for all priors π and gain functions g; equivalently, whenever D can be written as a post-processing of C (i.e., as $\mathrm{C} \,;\, \mathrm{R}$ for some channel R).

IV. Strategy-Based Formalisation of Leakage

The examples from Sections II-A and II-B show that the traditional definition of dynamic leakage (Definition 4) fails in significant ways. In particular, it might result in negative leakage even when it is clear that information has flowed (Example 1a); or it might lead to the conclusion that post-processing is safe, even though there are reasonable circumstances where this does not hold (Example 2a); it might also lead to the opposite conclusion: that post-processing is not safe, even though a more careful analysis proves this to be incorrect (we illustrate this case later in Section V).

As identified in Section II, these correspond to the properties of monotonicity and DPI, respectively. It turns out that, in general, monotonicity and DPI cannot be guaranteed when considering a dynamic perspective (as Example 2a indicates), but there are particular scenarios in which these axioms always hold. With this in mind, our goal is to provide a proper formalisation of dynamic leakage that satisfies these fundamental properties when possible, and that yields sound results when such properties are not guaranteed to hold. In addition, we require that our formalisation recovers the static perspective when taking into account all possible observations.

In what follows we present a novel formalisation of leakage which allows for both dynamic and static leakage analysis. We provide intuition for our measure in Section IV-A and

in Section IV-B we redefine g-vulnerability, ℓ-uncertainty and information leakage in terms of adversarial strategies guiding attackers' decisions. In Section IV-C we revisit the motivating examples from Section II, through the lens of our new formalisation. In Section IV-D we discuss the limitations of our formalisation, in that it requires a notion of ordering between states of knowledge. In Section IV-E we demonstrate desirable properties of the strategy-based formulation that hold under a dynamic perspective. Finally, in Section IV-F we show that our formalisation coincides with the traditional definitions adopted in QIF when considering a static perspective.

A. Defining leakage via actions: A thought experiment

In Example 1a the dynamic leakage computed using the traditional formalisation of Definition 4, with the attacker modelled by the loss function ℓ_H (8), produces a negative value even though we know that information flow has occurred. In that example the prior was $\{7/8, 1/16, 1/16\}$, and the computed posteriors were $p_{X|a} = \{1, 0, 0\}$ and $p_{X|b} = \{0, 1/2, 1/2\}$, occurring with probability $7/8$ and $1/8$, respectively. The first posterior (observation a) has all weight on secret x_0, meaning that the adversary is completely certain, and so it leads to a positive dynamic leakage of 0.67. In contrast, the second posterior (observation b), although less likely to be observed in practice, yields a negative dynamic leakage value of -0.33.

Recall from Section II-C that the issue in this case is that the set of states of knowledge considered in the analyses a priori and a posteriori are not compatible. To resolve this issue, consider the following **thought experiment**: imagine that there are two colluding adversaries: Bob, whose belief about the secret is formed before any observation is made (a prior), and Alice, whose belief results from a Bayesian update after observing a run of the system (a posterior), both of whom have the same objective (a shared gain function). Bob can compute his maximum expected gain (with respect to his belief) using the shared gain function; this results in a set of actions which are optimal from his point of view. Bob communicates his optimal actions to the more knowledgeable adversary, Alice, who fixes the posterior she computed as the baseline distribution on secret values according to which gain will be measured. She then has two options: gauge the gain using Bob's actions (chosen according to *his* belief) or using her actions (chosen according to *her* belief, i.e., the baseline).

The difference between the expected gain of Bob's actions versus Alice's actions, both computed against the most refined knowledge available as a baseline (Alice's posterior), represents the advantage of choosing optimal actions according to more refined beliefs versus choosing them according to less refined ones. It turns out that this difference is always non-negative (Bob's actions will never yield a higher vulnerability than Alice's when the baseline is Alice's posterior) and recovers static leakage under both expectation and max-case.

B. Strategy-based dynamic g-leakage

We now provide our definition for a dynamic leakage measure based on probability distributions over actions, which we term *strategies*. Our strategy-based leakage measure relies on our knowledge of the order in which information flow occurs (as when Alice has more knowledge than Bob in the thought experiment from Section IV-A), capturing the idea that information flow corresponds to the difference in gain obtained by choosing actions — i.e., defining a *strategy* — based on a more refined state of knowledge about the secret value than on a less refined one. The gain of such actions is then measured against the most refined state of knowledge about secrets available — i.e., the *baseline*. We start by defining an adversarial strategy, obtained from the adversary's belief:

Definition 8 (Adversarial strategy)**.** Given a gain function $g : \mathcal{W} \times \mathcal{X} \to \mathbb{R}$ modelling the adversary's goal and a probability distribution $q : \mathbb{D}\mathcal{X}$ representing the adversary's belief about the secret, an optimal strategy s is defined as a probability distribution over the actions $\mathcal{W}$ such that $s_{w^*} > 0$ only if w^* is an optimal action in the sense that it maximises the adversary's gain with respect to g and q; that is,

$$w^* \in \operatorname*{argmax}_{w \in \mathcal{W}} \sum_{x \in \mathcal{X}} q_x \, g(w, x)$$

Analogously, for loss functions, replace argmax with argmin:

$$w^* \in \operatorname*{argmin}_{w \in \mathcal{W}} \sum_{x \in \mathcal{X}} q_x \, \ell(w, x)$$

Remark 2. In this paper we assume that the set of optimal actions is finite (but not necessarily the set of actions $\mathcal{W}$).[5]

We note that there might be multiple strategies that are optimal according to the adversary's belief. We write $\mathrm{st}_g(q)$ for the *set* of all optimal strategies with respect to a belief q and gain function g. Then, taking $p : \mathbb{D}\mathcal{X}$ as the baseline, the g-vulnerability that results from any $s \in \mathrm{st}_g(q)$ can be defined as the expected gain of the optimal actions selected by s, when secrets are distributed according to the baseline p:

$$\sum_{w \in \mathcal{W}} s_w \sum_{x \in \mathcal{X}} p_x \, g(w, x) \tag{12}$$

The adversary is, in principle, allowed to pick any strategy in $\mathrm{st}_g(q)$. Nevertheless, we argue that, since all strategies in $\mathrm{st}_g(q)$ are equally optimal, the adversary has no reason to favour one over another. As such, it seems reasonable to define dynamic g-vulnerability (dually, ℓ-uncertainty) as an average over $s \in \mathrm{st}_g(q)$, assigning the same probability to each s. A caveat of this definition is that $|\mathrm{st}_g(q)|$ may be infinite. This is nonetheless solved by observing that the average over $\mathrm{st}_g(q)$ converges to the strategy-based g-vulnerability of a uniform strategy (i.e., uniform over all optimal actions), denoted $\mathrm{st}^u_g(q)$. The next example illustrates this phenomenon:

Example 3. Let Bob's belief be $q = \{1/3, 1/3, 1/3\}$, Alice's belief after observing a run of the system be $p = \{0, 1/2, 1/2\}$, and their goal be modelled by the identity gain function $\mathbf{1}$ (7). From Bob's point of view, all three secret values are optimal guesses, as per Definition 8 with $\mathcal{W} = \mathcal{X}$. The number of

[5] We believe this is reasonable, as this is true for every case we have studied.

optimal strategies available for Bob is infinite, but we can circumvent this by limiting the strategies that Bob can choose.

Let the number of decimal places n assigned by Bob to the probability of each action be 1. The set of strategies available to Bob is $\{\{1,0,0\},\{0.9,0.1,0\},\ldots\}$. There are 66 strategies in total. Among these, 11 assign probability 0 to x_0, and so, according to Alice's knowledge, lead to a vulnerability of $1/2$, as per (12). Then, there are 10 strategies assigning probability 0.1 to x_0, resulting in a vulnerability of $9/20$ each. The sum of the vulnerabilities is $(11 \cdot 1/2)+(10 \cdot 9/20)+\cdots+(1 \cdot 0) = 22$. The average over the 66 strategies — when picked at random — then yields a vulnerability of $22/66 = 1/3$, exactly the same vulnerability if Bob chooses the uniform strategy. □

The convergence seen in Example 3 occurrs with any $n \geq 0$. For more details, see Appendix A. We thus define *strategy-based* measurements as the expected case over optimal actions chosen based on the adversary's belief q, when they pick one of such actions at random,[6] and when secrets follow a more refined knowledge p, the baseline:

Definition 9 (Strategy-based measurements)**.** Let $q : \mathbb{D}\mathcal{X}$ be the adversary's belief and $p : \mathbb{D}\mathcal{X}$ be the baseline. Then, for any gain function $g : \mathcal{W} \times \mathcal{X} \to \mathbb{R}$, the *strategy-based g-vulnerability* caused by q with respect to the baseline p is

$$V_g^{\mathrm{st}}(p \,\|\, q) \stackrel{\text{def}}{=} \sum_{w \in \mathcal{W}} \mathrm{st}_g^u(q)_w \sum_{x \in \mathcal{X}} p_x \, g(w,x)$$

Dually, for any loss function $\ell : \mathcal{W} \times \mathcal{X} \to \mathbb{R}_{\geq 0} \cup \{\infty\}$,

$$U_\ell^{\mathrm{st}}(p \,\|\, q) \stackrel{\text{def}}{=} \sum_{w \in \mathcal{W}} \mathrm{st}_\ell^u(q)_w \sum_{x \in \mathcal{X}} p_x \, \ell(w,x)$$

Finally, we formalise the strategy-based dynamic leakage based on the difference between the vulnerability of the secret to the adversaries with more (posterior) and less (prior) refined beliefs. This measures the knowledge gain to the adversary after making an observation, and is defined (additively) as:

Definition 10 (Strategy-based dynamic leakage)**.** For any gain function $g : \mathcal{W} \times \mathcal{X} \to \mathbb{R}$, prior knowledge $\pi : \mathbb{D}\mathcal{X}$, channel $\mathrm{C} : \mathcal{X} \to \mathbb{D}\mathcal{Y}$ and output $y \in \mathcal{Y}$, the (additive) *strategy-based dynamic g-leakage* $\mathcal{L}_g^{\mathrm{st}}((\pi \triangleright \mathrm{C})_{X|y} \,\|\, \pi)$ is defined as

$$V_g^{\mathrm{st}}((\pi \triangleright \mathrm{C})_{X|y} \,\|\, (\pi \triangleright \mathrm{C})_{X|y}) - V_g^{\mathrm{st}}((\pi \triangleright \mathrm{C})_{X|y} \,\|\, \pi)$$

For loss functions, we define $\mathcal{L}_\ell^{\mathrm{st}}((\pi \triangleright \mathrm{C})_{X|y} \,\|\, \pi)$ as

$$U_\ell^{\mathrm{st}}((\pi \triangleright \mathrm{C})_{X|y} \,\|\, \pi) - U_\ell^{\mathrm{st}}((\pi \triangleright \mathrm{C})_{X|y} \,\|\, (\pi \triangleright \mathrm{C})_{X|y})$$

C. Motivating examples revisited

We now revisit the motivating examples from Section II and show how our definition of dynamic leakage resolves the identified issues of monotonicity and DPI. We start by revisiting Example 1a, to demonstrate that, from the point of view of an attacker, observing an output of a system should never reduce information, and thus a rational adversary would always rely on their posterior knowledge.

[6] Alternatively, we could consider the supremum over all strategies, assuming that the adversary always picks the best strategy. We argue that this might underestimate leakage, as there is nothing indicating to the adversary which strategy to choose. For instance, in Example 3 leakage would be 0.

Example 1b (continuing from page 2)**.** Recall the deterministic program that takes as input a 2-bit secret from the set of possible secrets $\mathcal{X} = \{00, 10, 11\}$, and outputs a if $X = 00$ or else b. This can be modelled by the deterministic channel B below, which, in conjunction with the adversary's prior knowledge $\pi = p_X$, gives the posterior knowledge $[\pi \triangleright \mathrm{B}]$:

$$\begin{array}{c} \pi \\ 00 \\ 10 \\ 11 \end{array} \begin{bmatrix} \frac{7}{8} \\ \frac{1}{16} \\ \frac{1}{16} \end{bmatrix} \begin{bmatrix}\triangleright\end{bmatrix} \begin{array}{c} \mathrm{B} \\ 00 \\ 10 \\ 11 \end{array} \begin{array}{cc} a & b \\ \end{array} \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 1 \end{bmatrix} = \begin{array}{c} [\pi \triangleright \mathrm{B}] \\ 00 \\ 10 \\ 11 \end{array} \begin{array}{cc} \frac{7}{8}\,a & \frac{1}{8}\,b \\ \end{array} \begin{bmatrix} 1 & 0 \\ 0 & \frac{1}{2} \\ 0 & \frac{1}{2} \end{bmatrix}$$

Now, the adversary's prior strategy with respect to the loss function ℓ_H (8) will be to guess that the real distribution from which the secret was drawn is exactly the prior distribution π. Therefore, the adversary's (uniform) strategy *a priori*, i.e., $\mathrm{st}_{\ell_H}^u(\pi)$, is a *point distribution* $[\pi]$, noting that in general

$$[x]_y \stackrel{\text{def}}{=} 1 \text{ if } x = y \text{ else } 0 \tag{13}$$

Then, program B was executed and produced an output b. We can compute the strategy-based ℓ_H-uncertainty of the prior π taking the posterior $(\pi \triangleright \mathrm{B})_{X|b}$ as the baseline to get

$$\begin{aligned} & U_{\ell_H}^{\mathrm{st}}((\pi \triangleright \mathrm{B})_{X|b} \,\|\, \pi) \\ = & \sum_{w \in \mathcal{W}} \mathrm{st}_{\ell_H}^u(\pi)_w \sum_{x \in \mathcal{X}} (\pi \triangleright \mathrm{B})_{x|b} \, \ell_H(w,x) && \text{\# Definition 9} \\ = & \sum_{x \in \mathcal{X}} (\pi \triangleright \mathrm{B})_{x|b} \, (-\log_2 \pi_x) && \begin{array}{l}\text{\# } \mathrm{st}_{\ell_H}^u(\pi) = [\pi]\text{, defined in (13);} \\ \text{\# Definition of } \ell_H \text{ in (8)}\end{array} \\ = & -\left(0 \log_2 \frac{7}{8} + \frac{1}{2} \log_2 \frac{1}{16} + \frac{1}{2} \log_2 \frac{1}{16}\right) = 4 \end{aligned}$$

which is greater than the strategy-based posterior uncertainty $U_{\ell_H}^{\mathrm{st}}((\pi \triangleright \mathrm{B})_{X|b} \,\|\, (\pi \triangleright \mathrm{B})_{X|b}) = U_{\ell_H}((\pi \triangleright \mathrm{B})_{X|b}) = 1$. Notice that this time leakage (Definition 10) is $4 - 1 = 3$, indicating that uncertainty decreased. This captures the fact that the adversary learned the first bit of the secret, and can be interpreted as the Kullback-Leibler (KL) divergence of the adversary's prior knowledge and the baseline, which we explain later.

To flesh out why our formulation works best in this example, recall that, in the traditional definition of dynamic leakage (Definition 4), the adversary's prior state of knowledge plays the role of both belief and baseline when measuring the prior vulnerability of the secret. This is not suited for a dynamic analysis, as it considers that when the adversary picked "00" as the most likely secret a priori, there was indeed a $7/8$ probability of this choice being the right one, but the more refined knowledge obtained after observing the execution of the system reveals that the secret "00" was never possible in the first place. Our measure captures that by assessing prior vulnerability as the success of the prior strategy, based on the adversary's prior belief, against the baseline (the posterior), which correctly reflects that the optimal prior choice was not going to be optimal in the light of the more refined baseline available for this run. Hence, the actual leakage is positive. □

Recall from Remark 1 that we defined uncertainty measures to be bounded. This would be trivially true if we had restricted loss functions to $\mathbb{R}_{\geq 0}$, but ℓ_H (8) has range $[0, \infty]$. Nevertheless, as the next result shows, we can confirm that the strategy-based ℓ_H-uncertainty satisfies such a requirement.

Lemma 1. *For any prior knowledge $\pi : \mathbb{D}\mathcal{X}$ and channel $\mathrm{C} : \mathcal{X} \to \mathbb{D}\mathcal{Y}$, it follows that both the strategy-based dynamic prior ℓ_H-uncertainty $U^{\mathrm{st}}_{\ell_H}((\pi \triangleright \mathrm{C})_{X|y} \,\|\, \pi)$ and the strategy-based dynamic posterior ℓ_H-uncertainty $U^{\mathrm{st}}_{\ell_H}((\pi \triangleright \mathrm{C})_{X|y} \,\|\, (\pi \triangleright \mathrm{C})_{X|y})$ are finite and non-negative.*

There is an interesting correspondence between the strategy-based ℓ_H-uncertainty and the Kullback-Leibler (KL) divergence of two distributions $p : \mathbb{D}\mathcal{X}$ and $q : \mathbb{D}\mathcal{X}$, defined as

$$D_{KL}(p \,\|\, q) \stackrel{\text{def}}{=} \sum_{x \in \mathcal{X}} p_x \log_2 \frac{p_x}{q_x} \tag{14}$$

Notice that the result obtained in Example 1b is exactly the KL divergence of the posterior $(\pi \triangleright \mathrm{C})_{X|b}$ and prior π:

$$\begin{aligned} & D_{KL}((\pi \triangleright \mathrm{C})_{X|b} \,\|\, \pi) \\ = \; & 0 \log_2 \frac{0}{7/8} + \frac{1}{2} \log_2 \frac{1/2}{1/16} + \frac{1}{2} \log_2 \frac{1/2}{1/16} = 3 \end{aligned}$$

In fact, as the next theorem states, this equivalence is not particular to Example 1b, and this is what motivated the use of the $\|$ operator in the notation for strategy-based measures.

Theorem 1. *Let $q : \mathbb{D}\mathcal{X}$ be the adversary's belief and $p : \mathbb{D}\mathcal{X}$ be the baseline. Then, $\mathcal{L}^{\mathrm{st}}_{\ell_H}(p \,\|\, q) = D_{KL}(p \,\|\, q)$.*

Notice that, in principle, KL divergence could be infinite. However, since we are not interested in generic distributions p and q but rather in a posterior distribution obtained from a prior and a channel, it turns out that strategy-based ℓ_H-leakage *is* bounded, and this follows immediately from Lemma 1.

Going back to Example 1b, Theorem 1 means that the value "3" we obtained as the information leakage, according to Definition 10, can be expressed in terms of the KL divergence of the prior and posterior states of knowledge. Therefore, it can be interpreted according to Shannon's source-coding theorem: it is the average number of extra bits needed to encode the secret when one believes that the secret was produced from a distribution π, but in reality it was obtained from the baseline $(\pi \triangleright \mathrm{C})_{X|b}$. From the attacker's point of view, this could be understood as the number of "yes/no questions" they save by changing from their prior to the posterior knowledge.[7]

Next, we revisit Example 2a, to show that the misleading conclusion of the preservation of the DPI, previously obtained with the traditional definition of dynamic vulnerability as in (3), is now solved under our formalisation. We note that this behaviour (a break in the DPI under a dynamic perspective, even though refinement holds in the static perspective) can only be observed in analyses where there is a third system that yields the baseline (in Example 2a, a system that reports the original answer, available to the data engineering team). Later, in Section IV-E, we show that the DPI always holds pairwise. This means that a single step of post-processing is safe, but (counter-intuitively) further steps might not be.

Example 2b (continuing from page 3)**.** The system P can be modelled by a channel $\mathrm{P} : \mathcal{X} \to \mathbb{D}\mathcal{X}$, where $\mathcal{X} = \{\text{NO}, \text{YES}\}$. Similarly, the second post-processing step S can be represented as a channel $\mathrm{S} : \mathcal{X} \to \mathbb{D}\mathcal{X}$. From these, we construct the new system P;S, as per Definition 1. These channels are as follows:

$$\begin{array}{c} \mathrm{P} \\ \text{NO} \\ \text{YES} \end{array} \begin{array}{c} \begin{array}{cc} \text{NO} & \text{YES} \end{array} \\ \begin{bmatrix} \frac{2}{3} & \frac{1}{3} \\ \frac{1}{3} & \frac{2}{3} \end{bmatrix} \end{array} ; \begin{array}{c} \mathrm{S} \\ \text{NO} \\ \text{YES} \end{array} \begin{array}{c} \begin{array}{cc} \text{NO} & \text{YES} \end{array} \\ \begin{bmatrix} 1 & 0 \\ \frac{1}{2} & \frac{1}{2} \end{bmatrix} \end{array} = \begin{array}{c} \mathrm{P};\mathrm{S} \\ \text{NO} \\ \text{YES} \end{array} \begin{array}{c} \begin{array}{cc} \text{NO} & \text{YES} \end{array} \\ \begin{bmatrix} \frac{5}{6} & \frac{1}{6} \\ \frac{2}{3} & \frac{1}{3} \end{bmatrix} \end{array}$$

Then, from channels P and P;S, and assuming a uniform prior $\upsilon = p_X$, we can construct $\mathscr{A}^{\mathrm{P}}$'s posterior knowledge $[\upsilon \triangleright \mathrm{P}]$ and $\mathscr{A}^{\mathrm{S}}$'s posterior knowledge $[\upsilon \triangleright \mathrm{P};\mathrm{S}]$ as, respectively,

$$\begin{array}{c} [\upsilon \triangleright \mathrm{P}] \\ \text{NO} \\ \text{YES} \end{array} \begin{array}{c} \begin{array}{cc} \frac{1}{2}\,\text{NO} & \frac{1}{2}\,\text{YES} \end{array} \\ \begin{bmatrix} \frac{2}{3} & \frac{1}{3} \\ \frac{1}{3} & \frac{2}{3} \end{bmatrix} \end{array} \quad \text{and} \quad \begin{array}{c} [\upsilon \triangleright \mathrm{P};\mathrm{S}] \\ \text{NO} \\ \text{YES} \end{array} \begin{array}{c} \begin{array}{cc} \frac{3}{4}\,\text{NO} & \frac{1}{4}\,\text{YES} \end{array} \\ \begin{bmatrix} \frac{5}{9} & \frac{1}{3} \\ \frac{4}{9} & \frac{2}{3} \end{bmatrix} \end{array}$$

Assuming that the secret answer was NO, the baseline we consider is that of the data engineering team, namely the point distribution [NO], against which we evaluate both $\mathscr{A}^{\mathrm{P}}$'s and $\mathscr{A}^{\mathrm{S}}$'s *a posteriori* strategies. Starting with $\mathscr{A}^{\mathrm{P}}$, recall that she observed YES. Therefore, her strategy w.r.t. the gain function $\mathbf{1}$ (7) is $\mathrm{st}^u_{\mathbf{1}}((\upsilon \triangleright \mathrm{P})_{X|\text{YES}}) = [\text{YES}] = \{0, 1\}$, meaning that she will guess YES. This gives a strategy-based vulnerability of

$$\begin{aligned} & V^{\mathrm{st}}_{\mathbf{1}}([\text{NO}] \,\|\, (\upsilon \triangleright \mathrm{P})_{X|\text{YES}}) \\ = \; & \sum_{w \in \mathcal{W}} \mathrm{st}^u_{\mathbf{1}}((\upsilon \triangleright \mathrm{P})_{X|\text{YES}})_w \sum_{x \in \mathcal{X}} [\text{NO}]_x \, \mathbf{1}(w, x) && \text{\# Definition 9} \\ = \; & \sum_{w \in \mathcal{W}} [\text{YES}]_w \sum_{x \in \mathcal{X}} [\text{NO}]_x \, \mathbf{1}(w, x) && \text{\# } \mathrm{st}^u_{\mathbf{1}}((\upsilon \triangleright \mathrm{P})_{X|\text{YES}}) = [\text{YES}] \\ = \; & \mathbf{1}(\text{YES}, \text{NO}) = 0 && \begin{array}{l} \text{\# Definition of [NO] and [YES] in (13);} \\ \text{\# Definition of } \mathbf{1} \text{ in (7)} \end{array} \end{aligned}$$

For adversary $\mathscr{A}^{\mathrm{S}}$, recall that he observed NO and thus his strategy according to his posterior knowledge $(\upsilon \triangleright \mathrm{P};\mathrm{S})_{X|\text{NO}}$ will be to guess that the original answer was NO. That is, $\mathrm{st}^u_{\mathbf{1}}((\upsilon \triangleright \mathrm{P};\mathrm{S})_{X|\text{NO}}) = [\text{NO}] = \{1, 0\}$. Following the same reasoning above, we get a vulnerability of

$$V^{\mathrm{st}}_{\mathbf{1}}([\text{NO}] \,\|\, (\upsilon \triangleright \mathrm{P};\mathrm{S})_{X|\text{NO}}) = \mathbf{1}(\text{NO}, \text{NO}) = 1$$

which is larger than the vulnerability computed for $\mathscr{A}^{\mathrm{P}}$. This shows that the strategy-based formalisation correctly leads to the conclusion that, from the perspective of the engineering team, who know the true answer, it is not always safe to replace P with P;S (albeit *on average* it is, as $\mathrm{P} \sqsubseteq \mathrm{P};\mathrm{S}$). □

D. Knowledge ordering and its limitations

Notice that Definition 9 of strategy-based g-vulnerability (or, equivalently, strategy-based ℓ-uncertainty) requires two probability distributions: $p \,\|\, q$. The distribution p is what we called the *baseline*, and it represents the knowledge of the more-informed adversary. However, we have not addressed how to identify who is the more-informed adversary.

[7] Although, from a privacy perspective, it is not immediately clear whether this kind of measure is well-suited, and prior work suggests otherwise [21].

We begin by noting that, given two distributions p and q, there is no information inherent to p or q that a rational adversary could rely on to decide which to use as a baseline. As we shall see later in Lemma 3, an adversary that derives their strategy from the baseline always reaches an optimal gain, implying that $\mathcal{L}^{\mathrm{st}}_g(p_A \,\|\, p_B) \geq 0$, but also $\mathcal{L}^{\mathrm{st}}_g(p_B \,\|\, p_A) \geq 0$. Hence, $\mathcal{L}^{\mathrm{st}}_g$ on its own is meaningless in terms of whether information flow has indeed occurred. This is not surprising, though, as this behaviour is observed in other well-known measures, such as the Kullback-Leibler divergence. In the case of $D_{\mathrm{KL}}(p \,\|\, q)$, p is viewed as the true probability distribution and q as the model, but looking at the distributions p and q without the context does not give any information, and flipping the roles of p and q still leads to a non-negative KL divergence.

However, when there is a refinement $\mathrm{C} \sqsubseteq \mathrm{D}$ (Definition 7), we know that, on average, an adversary $\mathscr{A}^{\mathrm{C}}$ who observes an output of channel C gains more information than an adversary $\mathscr{A}^{\mathrm{D}}$ with access only to D. Thus, although in the dynamic case it is possible that the strategy of $\mathscr{A}^{\mathrm{D}}$ is better than that of $\mathscr{A}^{\mathrm{C}}$, there is no way for either adversary to know this and, hence, there is no incentive for each to act differently than they do. As such, we argue that $\mathscr{A}^{\mathrm{C}}$ can be deemed more informed.

Remark 3. In the case where there is only one system under analysis, the more informed adversary is the one with the posterior knowledge, after observing an execution of the system (Alice, in the thought experiment). We note that this case can also be modelled as a refinement: given a prior π and a channel C, we can recover π (Bob's knowledge) as a post-processing of C by $\mathbb{1}$ via $[\pi \triangleright \mathrm{C}; \mathbb{1}] = [\pi \triangleright \mathbb{1}] = [\pi]$, where $\mathbb{1}$ is the channel that leaks nothing, i.e., a channel mapping every input to $\bot$, representing the lack of a useful observation.

From this reasoning, we argue that a knowledge ordering is *necessary* in order to draw sensible conclusions regarding dynamic leakage using our measure, and that $\sqsubseteq$ is an appropriate relation that can be used to define such an ordering. We remark that the refinement defined for average-case static measures is strictly stronger than refinement on max-case measures, and therefore constitutes the strongest order used in QIF [22]. Going back to the operation $p \,\|\, q$ in $V^{\mathrm{st}}_g(p \,\|\, q)$, there must be a prior knowledge $\pi : \mathbb{D}\mathcal{X}$, channels $\mathrm{B} : \mathcal{X} \to \mathbb{D}\mathcal{Y}$ and $\mathrm{C} : \mathcal{X} \to \mathbb{D}\mathcal{Z}$, and outputs $y \in \mathcal{Y}$ and $z \in \mathcal{Z}$ such that

- $\mathrm{B} \sqsubseteq \mathrm{C}$, and thus there exists R such that $\mathrm{C} = \mathrm{B}\,;\mathrm{R}$;
- $p = (\pi \triangleright \mathrm{B})_{X|y}$ and $q = (\pi \triangleright \mathrm{C})_{X|z}$; and
- $\mathrm{R}_{y,z} > 0$, so that z is the result of post-processing y.

E. Towards an axiomatisation of dynamic leakage

We now turn our attention to three important axioms that hold under a dynamic perspective, when comparing two adversaries that satisfy the knowledge ordering discussed in Section IV-D: single-step monotonicity, single-step data-processing inequality and non-interference. To show them, we first consider the following lemma that states that the strategy-based g-vulnerability of $p \,\|\, p$ (i.e., when the baseline is p itself) is just the traditional g-vulnerability of the distribution p:

Lemma 2. *Let $p : \mathbb{D}\mathcal{X}$ be both the adversary's belief and the baseline. Then, for any gain function $g : \mathcal{W} \times \mathcal{X} \to \mathbb{R}$,*

$$V^{\mathrm{st}}_g(p \,\|\, p) = V_g(p)$$

Dually, for any loss function $\ell : \mathcal{W} \times \mathcal{X} \to \mathbb{R}_{\geq 0} \cup \{\infty\}$,

$$U^{\mathrm{st}}_\ell(p \,\|\, p) = U_\ell(p)$$

Moreover, vulnerability is maximum when the belief from which actions are chosen coincides with the baseline. This means that, when evaluating two adversaries with beliefs represented by distributions p and q, the result is only meaningful if one knows that p and q satisfy the knowledge ordering in some direction, so that there is a well-defined baseline.

Lemma 3. *Let $q : \mathbb{D}\mathcal{X}$ be the adversary's belief and $p : \mathbb{D}\mathcal{X}$ be the baseline. Then, for any gain function $g : \mathcal{W} \times \mathcal{X} \to \mathbb{R}$,*

$$V^{\mathrm{st}}_g(p \,\|\, q) \leq V^{\mathrm{st}}_g(p \,\|\, p) = V_g(p)$$

with equality holding when $q = p$. Dually, for any loss function $\ell : \mathcal{W} \times \mathcal{X} \to \mathbb{R}_{\geq 0} \cup \{\infty\}$, it follows that

$$U^{\mathrm{st}}_\ell(p \,\|\, q) \geq U^{\mathrm{st}}_\ell(p \,\|\, p) = U_\ell(p)$$

We are now ready to state the three axioms of single-step monotonicity, single-step DPI and non-interference, under a dynamic perspective. We begin with single-step monotonicity, which corresponds to the perspective of an attacker, who starts with a prior knowledge π, knows the implementation details of the system C under analysis, and can observe the result of an execution of system C, but cannot observe any intermediate value flowing through the sequential steps that constitute C:

Theorem 2 (Single-step monotonicity)**.** *Let $\pi : \mathbb{D}\mathcal{X}$ be the adversary's prior belief, and consider a channel $\mathrm{C} : \mathcal{X} \to \mathbb{D}\mathcal{Y}$ producing $y \in \mathcal{Y}$. Then, for any gain function $g : \mathcal{W} \times \mathcal{X} \to \mathbb{R}$,*

$$V^{\mathrm{st}}_g((\pi \triangleright \mathrm{C})_{X|y} \,\|\, \pi) \leq V^{\mathrm{st}}_g((\pi \triangleright \mathrm{C})_{X|y} \,\|\, (\pi \triangleright \mathrm{C})_{X|y})$$

Dually, for any loss function $\ell : \mathcal{W} \times \mathcal{X} \to \mathbb{R}_{\geq 0} \cup \{\infty\}$,

$$U^{\mathrm{st}}_\ell((\pi \triangleright \mathrm{C})_{X|y} \,\|\, \pi) \geq U^{\mathrm{st}}_\ell((\pi \triangleright \mathrm{C})_{X|y} \,\|\, (\pi \triangleright \mathrm{C})_{X|y})$$

The (single-step) data-processing inequality axiom for the dynamic perspective states that, whenever comparing two adversaries with access to channels C and D, whose states of knowledge are known to satisfy the knowledge ordering — i.e., $\mathrm{C} \sqsubseteq \mathrm{D}$ — it is *rational* to replace channel C with channel D. Equivalently, a rational adversary would always prefer to have the extra knowledge that comes with channel C:

Theorem 3 (Single-step DPI)**.** *Let $\pi : \mathbb{D}\mathcal{X}$ be the adversary's prior belief, and consider two channels $\mathrm{C} : \mathcal{X} \to \mathbb{D}\mathcal{Y}$ and $\mathrm{D} : \mathcal{X} \to \mathbb{D}\mathcal{Z}$ such that $\mathrm{C} \sqsubseteq \mathrm{D}$. Let $y \in \mathcal{Y}$ and $z \in \mathcal{Z}$ such that $\mathrm{R}_{y,z} > 0$. Then, for any gain function $g : \mathcal{W} \times \mathcal{X} \to \mathbb{R}$,*

$$V^{\mathrm{st}}_g((\pi \triangleright \mathrm{C})_{X|y} \,\|\, (\pi \triangleright \mathrm{D})_{X|z}) \leq V^{\mathrm{st}}_g((\pi \triangleright \mathrm{C})_{X|y} \,\|\, (\pi \triangleright \mathrm{C})_{X|y})$$

Dually, for any loss function $\ell : \mathcal{W} \times \mathcal{X} \to \mathbb{R}_{\geq 0} \cup \{\infty\}$,

$$U^{\mathrm{st}}_\ell((\pi \triangleright \mathrm{C})_{X|y} \,\|\, (\pi \triangleright \mathrm{D})_{X|z}) \geq U^{\mathrm{st}}_\ell((\pi \triangleright \mathrm{C})_{X|y} \,\|\, (\pi \triangleright \mathrm{C})_{X|y})$$

Recall from Remark 3 that monotonicity can be viewed as a particular case of DPI, by setting channel $\mathrm{D} = \mathbb{1}$. With this in mind, notice that both single-step monotonicity and single-step DPI require that the baseline is exactly one of the posteriors obtained from channel C, with $\mathrm{C} \sqsubseteq \mathrm{D}$. In contrast, whenever the baseline comes from a third system $\mathrm{B} : \mathcal{X} \to \mathbb{D}\mathcal{O}$, with $\mathrm{B} \sqsubseteq \mathrm{C} \sqsubseteq \mathrm{D}$, $\mathrm{B} \neq \mathrm{C}$ and $\mathrm{B} \neq \mathrm{D}$ (as in Example 2a), the DPI is not necessarily satisfied; that is, it might be that

$$V_g^{\mathrm{st}}((\pi \triangleright \mathrm{B})_{X|o} \,\|\, (\pi \triangleright \mathrm{D})_{X|z}) > V_g^{\mathrm{st}}((\pi \triangleright \mathrm{B})_{X|o} \,\|\, (\pi \triangleright \mathrm{C})_{X|y})$$

We call this a *multi-step* analysis, as there are (at least) two dynamic steps: $o \to y$ and then $y \to z$. We will return to multi-step analysis in Section VI, to understand its implications.

Finally, the non-interference axiom states that whenever the system corresponds to channel $\mathbb{1}$, i.e., the channel that leaks nothing, then no information leakage can occur. This is true under a static perspective, as $V_g[\pi \triangleright \mathbb{1}] = V_g[\pi] = V_g(\pi)$, and it is trivial to see under a dynamic perspective, as there is only one possible output $\bot$ from which the adversary learns nothing useful that could guide them in changing their strategy:

Theorem 4 (Non-interference)**.** *Let $\pi : \mathbb{D}\mathcal{X}$ be the adversary's prior belief. Then, for any gain function $g : \mathcal{W} \times \mathcal{X} \to \mathbb{R}$,*

$$V_g^{\mathrm{st}}((\pi \triangleright \mathbb{1})_{X|\bot} \,\|\, (\pi \triangleright \mathbb{1})_{X|\bot}) = V_g^{\mathrm{st}}((\pi \triangleright \mathbb{1})_{X|\bot} \,\|\, \pi)$$

Dually, for any loss function $\ell : \mathcal{W} \times \mathcal{X} \to \mathbb{R}_{\geq 0} \cup \{\infty\}$,

$$U_\ell^{\mathrm{st}}((\pi \triangleright \mathbb{1})_{X|\bot} \,\|\, (\pi \triangleright \mathbb{1})_{X|\bot}) = U_\ell^{\mathrm{st}}((\pi \triangleright \mathbb{1})_{X|\bot} \,\|\, \pi)$$

F. Consistency with the static perspective

In addition to the properties verified in Section IV-E under a dynamic perspective, we require the strategy-based formalisation to be consistent with the results obtained under a static perspective. We begin by showing that the proposed formalisation of dynamic g-vulnerability (dually, ℓ-uncertainty) recovers an important property of the traditional definition: the expected posterior g-vulnerability is the expectation of the strategy-based dynamic posterior g-vulnerability over all posteriors:

Theorem 5. *Let $\pi : \mathbb{D}\mathcal{X}$ be the adversary's prior belief, and consider a channel $\mathrm{C} : \mathcal{X} \to \mathbb{D}\mathcal{Y}$ modelling the system of interest. Then, for any gain function $g : \mathcal{W} \times \mathcal{X} \to \mathbb{R}$,*

$$\sum_{y \in \mathcal{Y}} (\pi \triangleright \mathrm{C})_y \, V_g^{\mathrm{st}}((\pi \triangleright \mathrm{C})_{X|y} \,\|\, (\pi \triangleright \mathrm{C})_{X|y}) = V_g[\pi \triangleright \mathrm{C}]$$

Dually, for any loss function $\ell : \mathcal{W} \times \mathcal{X} \to \mathbb{R}_{\geq 0} \cup \{\infty\}$,

$$\sum_{y \in \mathcal{Y}} (\pi \triangleright \mathrm{C})_y \, U_\ell^{\mathrm{st}}((\pi \triangleright \mathrm{C})_{X|y} \,\|\, (\pi \triangleright \mathrm{C})_{X|y}) = U_\ell[\pi \triangleright \mathrm{C}]$$

An analagous property can be derived for the max-case (or min-case, when computing ℓ-uncertainty). That is, max-case posterior g-vulnerability is the maximum strategy-based dynamic posterior g-vulnerability over all posteriors:

Theorem 6. *Let $\pi : \mathbb{D}\mathcal{X}$ be the adversary's prior belief, and consider a channel $\mathrm{C} : \mathcal{X} \to \mathbb{D}\mathcal{Y}$ modelling the system of interest. Then, for any gain function $g : \mathcal{W} \times \mathcal{X} \to \mathbb{R}$,*

$$\max_{y \in \mathcal{Y}} V_g^{\mathrm{st}}((\pi \triangleright \mathrm{C})_{X|y} \,\|\, (\pi \triangleright \mathrm{C})_{X|y}) = V_g^{\max}[\pi \triangleright \mathrm{C}]$$

Dually, for any loss function $\ell : \mathcal{W} \times \mathcal{X} \to \mathbb{R}_{\geq 0} \cup \{\infty\}$,

$$\min_{y \in \mathcal{Y}} U_\ell^{\mathrm{st}}((\pi \triangleright \mathrm{C})_{X|y} \,\|\, (\pi \triangleright \mathrm{C})_{X|y}) = U_\ell^{\min}[\pi \triangleright \mathrm{C}]$$

Theorems 5 and 6 follow directly from Lemma 2, from which we get that $V_g^{\mathrm{st}}((\pi \triangleright \mathrm{C})_{X|y} \,\|\, (\pi \triangleright \mathrm{C})_{X|y})$ is exactly the traditional $V_g((\pi \triangleright \mathrm{C})_{X|y})$. A perhaps more interesting property is that we can reconstruct the traditional prior g-vulnerability by averaging the strategy-based dynamic prior g-vulnerability over each possible posterior. To illustrate, recall the thought experiment from Section IV-A. If Alice has more information than Bob (because she made an observation that Bob has not), but she decides to follow Bob's strategy, she will end up with the same vulnerability as Bob would have computed. In the end this shows that knowledge does not give you an advantage, unless it guides the design of strategies.

Theorem 7. *Let $\pi : \mathbb{D}\mathcal{X}$ be the adversary's prior belief, and consider a channel $\mathrm{C} : \mathcal{X} \to \mathbb{D}\mathcal{Y}$ modelling the system of interest. Then, for any gain function $g : \mathcal{W} \times \mathcal{X} \to \mathbb{R}$,*

$$\sum_{y \in \mathcal{Y}} (\pi \triangleright \mathrm{C})_y \, V_g^{\mathrm{st}}((\pi \triangleright \mathrm{C})_{X|y} \,\|\, \pi) = V_g(\pi)$$

Dually, for any loss function $\ell : \mathcal{W} \times \mathcal{X} \to \mathbb{R}_{\geq 0} \cup \{\infty\}$,

$$\sum_{y \in \mathcal{Y}} (\pi \triangleright \mathrm{C})_y \, U_\ell^{\mathrm{st}}((\pi \triangleright \mathrm{C})_{X|y} \,\|\, \pi) = U_\ell(\pi)$$

Alice's extra knowledge can come from two sources: either she observed something that Bob has not or Bob's observation was the result of post-processing the output observed by Alice.[8] The next two theorems thus show that the extra knowledge due to the lack of post-processing is, again, only useful if it guides the construction of strategies:

Theorem 8. *Let $\pi : \mathbb{D}\mathcal{X}$ be the adversary's prior belief, and consider two channels $\mathrm{C} : \mathcal{X} \to \mathbb{D}\mathcal{Y}$ and $\mathrm{D} : \mathcal{X} \to \mathbb{D}\mathcal{Z}$, $\mathrm{C} \sqsubseteq \mathrm{D}$. For any gain function $g : \mathcal{W} \times \mathcal{X} \to \mathbb{R}$, the* expected *posterior g-vulnerability $V_g[\pi \triangleright \mathrm{D}]$ is recovered as*

$$\sum_{z \in \mathcal{Z}} \sum_{y \in \mathcal{Y}} (\pi \triangleright \mathrm{C})_y \, \mathrm{R}_{y,z} \, V_g^{\mathrm{st}}((\pi \triangleright \mathrm{C})_{X|y} \,\|\, (\pi \triangleright \mathrm{D})_{X|z})$$

Dually, for any loss function $\ell : \mathcal{W} \times \mathcal{X} \to \mathbb{R}_{\geq 0} \cup \{\infty\}$, the expected *posterior ℓ-uncertainty $U_\ell[\pi \triangleright \mathrm{D}]$ is recovered as*

$$\sum_{z \in \mathcal{Z}} \sum_{y \in \mathcal{Y}} (\pi \triangleright \mathrm{C})_y \, \mathrm{R}_{y,z} \, U_\ell^{\mathrm{st}}((\pi \triangleright \mathrm{C})_{X|y} \,\|\, (\pi \triangleright \mathrm{D})_{X|z})$$

Theorem 9. *Let $\pi : \mathbb{D}\mathcal{X}$ be the adversary's prior belief, and consider two channels $\mathrm{C} : \mathcal{X} \to \mathbb{D}\mathcal{Y}$ and $\mathrm{D} : \mathcal{X} \to \mathbb{D}\mathcal{Z}$, $\mathrm{C} \sqsubseteq \mathrm{D}$. For any gain function $g : \mathcal{W} \times \mathcal{X} \to \mathbb{R}$, the* max-case *posterior g-vulnerability $V_g^{\max}[\pi \triangleright \mathrm{D}]$ is recovered as*

$$\max_{z \in \mathcal{Z}} \frac{1}{(\pi \triangleright \mathrm{D})_z} \sum_{y \in \mathcal{Y}} (\pi \triangleright \mathrm{C})_y \, \mathrm{R}_{y,z} \, V_g^{\mathrm{st}}((\pi \triangleright \mathrm{C})_{X|y} \,\|\, (\pi \triangleright \mathrm{D})_{X|z})$$

[8] If the output observed by Bob was not the result of post-processing Alice's observation, then Alice's and Bob's states of knowledge are not necessarily comparable. The measurement of dynamic leakage relies on the knowledge ordering (Section IV-D), which is only known from the context; there is nothing explicit in the distributions of Alice and Bob which can always be used to determine who has more knowledge than the other.

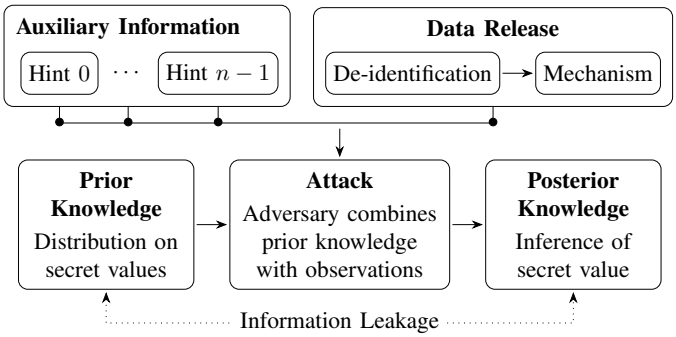


Fig. 3: Attack model against privacy-preserving data releases.

Dually, for any loss function $\ell : \mathcal{W} \times \mathcal{X} \to \mathbb{R}_{\geq 0} \cup \{\infty\}$*, the* min-case *posterior* ℓ*-uncertainty* $U_\ell^{\min}[\pi \triangleright \mathrm{D}]$ *is recovered as*

$$\min_{z \in \mathcal{Z}} \frac{1}{(\pi \triangleright \mathrm{D})_z} \sum_{y \in \mathcal{Y}} (\pi \triangleright \mathrm{C})_y \, \mathrm{R}_{y,z} \, U_\ell^{\mathrm{st}}((\pi \triangleright \mathrm{C})_{X|y} \,\|\, (\pi \triangleright \mathrm{D})_{X|z})$$

V. Case study: privacy-preserving data releases

In this section we show a dynamic scenario in which post-processing is safe, but the traditional definition of dynamic vulnerability (Definition 3) fails to capture that. This example is set in the context of privacy-preserving dataset releases.

To satisfy privacy constraints, datasets may undergo various transformations, which we abstract into two major steps. First, ***de-identification***, in which obvious personal identifiers, such as names or official ids, are stripped from the original data. This is modelled as a channel $\mathrm{D} : \mathcal{X} \to \mathbb{D}\mathcal{Y}$ that maps the original datasets to de-identified datasets. Second, mitigation, where one or more *privacy* ***mechanisms*** are employed to protect against different kinds of privacy attacks. This is modelled as a channel $\mathrm{M} : \mathcal{Y} \to \mathbb{D}\mathcal{Z}$ that maps de-identified datasets to sanitised datasets. The full privacy-preserving data-release process can be described by the cascading $\mathrm{D} \,;\, \mathrm{M} : \mathcal{X} \to \mathbb{D}\mathcal{Z}$ (Definition 1), depicted as "Data Release" in Figure 3.

We assume that the adversary has a prior knowledge about the original datasets. Then, once the data is actually published, the adversary observes a sanitised dataset, which leads them to update their knowledge. We assume also that the adversary has access to some auxiliary information that can be used to further improve their knowledge, often called *quasi-identifiers* (QIDs) [23, 24, 25, 26]. Formally, the adversary observes (i) the sanitised dataset produced by the data-release system $\mathrm{D} \,;\, \mathrm{M}$; and (ii) some QID values of one of the individuals, which can be viewed as the output of another channel $\mathrm{H} : \mathcal{X} \to \mathbb{D}\mathcal{Q}$ that produces a ***hint*** from that user's record.

We allow multiple hints (see Figure 3) and assume these are independent of each other, and also of the output of $\mathrm{D};\mathrm{M}$. This translates to the parallel composition $\mathrm{H}^0 \mathbin{⦶} \cdots \mathbin{⦶} \mathrm{H}^{n-1} \mathbin{⦶} (\mathrm{D};\mathrm{M})$, defined for two channels $\mathrm{C} : \mathcal{X} \to \mathbb{D}\mathcal{Y}$ and $\mathrm{D} : \mathcal{X} \to \mathbb{D}\mathcal{Z}$ as

$$(\mathrm{C} \mathbin{⦶} \mathrm{D})_{x,\langle y,z\rangle} \stackrel{\text{def}}{=} \mathrm{C}_{x,y} \, \mathrm{D}_{x,z} \quad \forall x \in \mathcal{Y} \text{ and } z \in \mathcal{Z} \qquad (15)$$

Figure 3 then describes the attack model for data releases.

Example 4. Consider an adversary $\mathscr{A}^{\mathrm{S}}$ who wants to determine which locations users visited, and how often, during a time period. Beforehand, he does not know anything. This is represented as a uniform prior over datasets, which are mappings from users to a multiset of locations visited by them:

$$\begin{array}{r} \upsilon \\ x_0 = \{\text{Alex}\colon \{\text{A, A}\}, \text{Bob}\colon \{\text{B}\}\} \\ x_1 = \{\text{Alex}\colon \{\text{B}\}, \text{Bob}\colon \{\text{A, A}\}\} \\ x_2 = \{\text{Alex}\colon \{\text{A, B}\}, \text{Bob}\colon \{\text{A}\}\} \\ x_3 = \{\text{Alex}\colon \{\text{A}\}, \text{Bob}\colon \{\text{A, B}\}\} \\ \vdots \end{array} \begin{bmatrix} \frac{1}{|\mathcal{X}|} \\ \frac{1}{|\mathcal{X}|} \\ \frac{1}{|\mathcal{X}|} \\ \frac{1}{|\mathcal{X}|} \\ \vdots \end{bmatrix}$$

Now, suppose that $\mathscr{A}^{\mathrm{S}}$ gains access to the de-identified database $s = \{0\colon \{\text{A, A}\}, 1\colon \{\text{B}\}\}$. Moreover, he knows that the database he received was perturbed using the mechanism

> if location $=$ A then replace with (A $\oplus_{3/4}$ B)
> else replace with (A $\oplus_{1/4}$ B).

This means that the probability of observing the sanitised dataset s given, e.g., that the secret is x_2, is $3/4 \cdot 1/4 \cdot 1/4 = 3/64$, noting that the mechanism is applied onto each location in the dataset. As such, when lifted to the domain of datasets, the mechanism can be represented by the following channel:

$$\begin{array}{r} \mathrm{M} \\ d_0 = \{1\colon \{\text{A, A}\}, 2\colon \{\text{B}\}\} \\ d_1 = \{1\colon \{\text{A, B}\}, 2\colon \{\text{A}\}\} \\ \vdots \end{array} \begin{array}{c} s \quad \cdots \\ \begin{bmatrix} \frac{27}{64} & \cdots \\ \frac{3}{64} & \cdots \\ \vdots & \vdots \end{bmatrix} \end{array}$$

where $s = \{0\colon \{\text{A, A}\}, 1\colon \{\text{B}\}\}$ is the sanitised dataset observed by $\mathscr{A}^{\mathrm{S}}$ and d_i is one of the possible de-identified datasets produced by channel D. The system $\mathrm{D} \,;\, \mathrm{M}$ is then

$$\begin{array}{c} \mathrm{D} \\ x_0 \\ x_1 \\ x_2 \\ x_3 \\ \vdots \end{array} \begin{array}{c} d_0 \quad d_1 \quad \cdots \\ \begin{bmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ \vdots & \vdots & \vdots \end{bmatrix} \end{array} \begin{array}{c} \mathrm{M} \\ d_0 \\ d_1 \\ \vdots \end{array} \begin{array}{c} s \quad \cdots \\ \begin{bmatrix} \frac{27}{64} & \cdots \\ \frac{3}{64} & \cdots \\ \vdots & \vdots \end{bmatrix} \end{array} = \begin{array}{c} \mathrm{D} \,;\, \mathrm{M} \\ x_0 \\ x_1 \\ x_2 \\ x_3 \\ \vdots \end{array} \begin{array}{c} s \quad \cdots \\ \begin{bmatrix} \frac{27}{64} & \cdots \\ \frac{27}{64} & \cdots \\ \frac{3}{64} & \cdots \\ \frac{3}{64} & \cdots \\ \vdots & \vdots \end{bmatrix} \end{array}$$

Suppose also that $\mathscr{A}^{\mathrm{S}}$ observes the hint that Alex visited location A, and a histogram $k = \{\text{A}\colon 2, \text{B}\colon 1\}$ with the number of occurrences of each location over the entire dataset. Let $\mathrm{H}^{\mathrm{Hist}} : \mathcal{X} \to \mathbb{D}\mathcal{K}$ model the system that maps datasets to histograms, and let $\mathrm{H}^{\mathrm{Alex}}$ reveal one of Alex's locations, defined in terms of the frequency of locations in her record. For example, $\mathrm{H}^{\mathrm{Alex}}_{x_2,\mathrm{A}} = 1/2$, as location A corresponds to half of the locations visited by Alex in x_2. Then, $\mathscr{A}^{\mathrm{S}}$ has the composition

$$\begin{array}{c} \mathrm{H}^{\mathrm{Hist}} \\ x_0 \\ x_1 \\ x_2 \\ x_3 \\ \vdots \end{array} \begin{array}{c} k \quad \cdots \\ \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 0 & \vdots \end{bmatrix} \end{array} \mathbin{⦶} \begin{array}{c} \mathrm{H}^{\mathrm{Alex}} \\ x_0 \\ x_1 \\ x_2 \\ x_3 \\ \vdots \end{array} \begin{array}{c} \text{A} \quad \text{B} \\ \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ \frac{1}{2} & \frac{1}{2} \\ 1 & 0 \\ \vdots & \vdots \end{bmatrix} \end{array} = \begin{array}{c} \langle k, \text{A}\rangle \quad \langle k, \text{B}\rangle \quad \cdots \\ \begin{bmatrix} 1 & 0 & \cdots \\ 0 & 1 & \cdots \\ \frac{1}{2} & \frac{1}{2} & \cdots \\ 1 & 0 & \cdots \\ 0 & 0 & \vdots \end{bmatrix} \end{array}$$

Let $\mathrm{S} = \mathrm{H}^{\mathrm{Hist}} \circledcirc \mathrm{H}^{\mathrm{Alex}} \circledcirc (\mathrm{D}\,;\mathrm{M})$. By combining his uniform prior υ with the two components above, $\mathscr{A}^{\mathrm{S}}$ can construct the posterior knowledge $(\upsilon \triangleright \mathrm{S})_{X|\langle k,\mathrm{A},s\rangle} = \{{}^{6}\!/_{7}, 0, {}^{1}\!/_{21}, {}^{2}\!/_{21}, \ldots\}$. From this, given the observation $\langle k, \mathrm{A}, s\rangle$, we can conclude that $\mathscr{A}^{\mathrm{S}}$'s best action is to guess x_0, which, following the traditional definition of dynamic Bayes vulnerability (Definition 3), gives him a chance of success of $V_{\mathbf{1}}((\upsilon \triangleright \mathrm{S})_{X|\langle k,\mathrm{A},s\rangle}) = {}^{6}\!/_{7}$.

Now, consider another adversary $\mathscr{A}^{\mathrm{P}}$ who observes the original (de-identified) data $d_1 = \{0\colon \{\mathrm{A}, \mathrm{B}\}, 1\colon \{\mathrm{A}\}\}$ before being sanitised onto dataset s, in addition to the histogram k (which can be implicitly derived from d_1) and the hint that Alex visited location A. Then, she can construct the following posterior knowledge, where $\mathrm{P} = \mathrm{H}^{\mathrm{Hist}} \circledcirc \mathrm{H}^{\mathrm{Alex}} \circledcirc \mathrm{D}$: $(\upsilon \triangleright \mathrm{P})_{X|\langle k,\mathrm{A},d_1\rangle} = \{0, 0, {}^{1}\!/_{3}, {}^{2}\!/_{3}, \ldots\}$. Therefore, given the observation $\langle k, \mathrm{A}, d_1\rangle$, $\mathscr{A}^{\mathrm{P}}$'s best action is to guess x_3, which gives her a chance of success of $V_{\mathbf{1}}((\upsilon \triangleright \mathrm{P})_{X|\langle k,\mathrm{A},d_1\rangle}) = {}^{2}\!/_{3}$.

Notice that $V_{\mathbf{1}}((\upsilon \triangleright \mathrm{S})_{X|\langle k,\mathrm{A},s\rangle}) > V_{\mathbf{1}}((\upsilon \triangleright \mathrm{P})_{X|\langle k,\mathrm{A},d_1\rangle})$, suggesting that $\mathscr{A}^{\mathrm{S}}$ has a better chance of guessing the correct secret. Nonetheless, his posterior knowledge is skewed towards the dataset x_0, which contains the wrong data: we assumed that $\mathscr{A}^{\mathrm{P}}$ observed the real data, and there is no record $\{\mathrm{A}, \mathrm{A}\}$ in x_3. Hence, $\mathscr{A}^{\mathrm{S}}$ would actually make a wrong guess; that is, his chance of success *in practice* is zero.

Now, let us see what happens when we employ the proposed strategy-based formalisation. Formally, $\mathscr{A}^{\mathrm{S}}$'s strategy with respect to gain function $\mathbf{1}$ is a point distribution $[x_0]$; i.e., $\mathrm{st}^{u}_{\mathbf{1}}((\upsilon \triangleright \mathrm{S})_{X|\langle k,\mathrm{A},s\rangle}) = [x_0] = \{1, 0, 0, 0, \ldots\}$. In contrast, the strategy of $\mathscr{A}^{\mathrm{P}}$ is a point distribution $[x_3]$; i.e., $\mathrm{st}^{u}_{\mathbf{1}}((\upsilon \triangleright \mathrm{P})_{X|\langle k,\mathrm{A},d_1\rangle}) = [x_3] = \{0, 0, 0, 1, \ldots\}$. Given the two strategies above, and taking the posterior knowledge of adversary $\mathscr{A}^{\mathrm{P}}$ as the baseline for the evaluation of such strategies ($\mathrm{D} \sqsubseteq \mathrm{D}\,;\mathrm{M}$ implies that $\mathrm{P} \sqsubseteq \mathrm{S}$ [27, Thm 5]), we compute that $\mathscr{A}^{\mathrm{S}}$'s chance of determining the secret dataset is

$$
\begin{aligned}
& V^{\mathrm{st}}_{\mathbf{1}}((\upsilon \triangleright \mathrm{P})_{X|\langle k,\mathrm{A},d_1\rangle} \,\|(\upsilon \triangleright \mathrm{S})_{X|\langle k,\mathrm{A},s\rangle}) \\
= & \sum_{w\in\mathcal{W}} \mathrm{st}^{u}_{\mathbf{1}}((\upsilon \triangleright \mathrm{S})_{X|\langle k,\mathrm{A},s\rangle})_w \sum_{x\in\mathcal{X}} (\upsilon \triangleright \mathrm{P})_{x|\langle k,\mathrm{A},d_1\rangle}\, \mathbf{1}(w,x) \\
= & \sum_{w\in\mathcal{W}} [x_0] \sum_{x\in\mathcal{X}} (\upsilon \triangleright \mathrm{P})_{x|\langle k,\mathrm{A},d_1\rangle}\, \mathbf{1}(w,x) \\
= & \ (\upsilon \triangleright \mathrm{P})_{x_0|\langle k,\mathrm{A},d_1\rangle} = 0 \qquad \begin{array}{l}\text{\# Definition of } [x_0] \text{ in (13);}\\ \text{\# Definition of } \mathbf{1} \text{ in (7)}\end{array}
\end{aligned}
$$

which is smaller than $\mathscr{A}^{\mathrm{P}}$'s chance of success, as expected:

$$
\begin{aligned}
& V^{\mathrm{st}}_{\mathbf{1}}((\upsilon \triangleright \mathrm{P})_{X|\langle k,\mathrm{A},d_1\rangle} \,\|(\upsilon \triangleright \mathrm{P})_{X|\langle k,\mathrm{A},d_1\rangle}) \\
= & \ V_{\mathbf{1}}((\upsilon \triangleright \mathrm{P})_{X|\langle k,\mathrm{A},d_1\rangle}) = {}^{2}\!/_{3} \qquad \square
\end{aligned}
$$

As with Example 2a, the example above shows how the traditional definition of information-flow measures leads to incorrect conclusions when comparing specific executions of systems (that is, dynamic scenarios). And, as the example illustrates, this issue is solved by our formalisation.

VI. Multi-step analysis

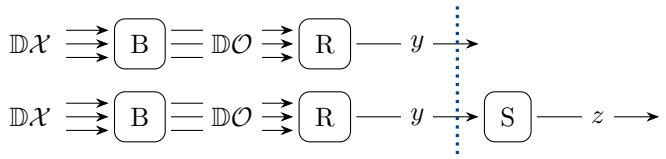
In Section IV-E we proved that the strategy-based formalisation of dynamic leakage satisfies relaxed versions of the axioms of monotonicity and of the data-processing inequality. More specifically, these properties are satisfied whenever the analysis involves a single $\mathrm{input} \to \mathrm{output}$ dynamic step. Figure 4a depicts this kind of analysis: the upper diagram represents a dynamic analysis with respect to some output $y \in \mathcal{Y}$ from channel $\mathrm{C} = \mathrm{B}\,;\mathrm{R}$, obtained from a sequence of steps $\mathbb{D}\mathcal{X} \to \mathbb{D}\mathcal{O} \to y$, whereas the lower diagram represents a dynamic analysis of some output $z \in \mathcal{Z}$ from channel $\mathrm{D} = \mathrm{C}\,;\mathrm{S}$, obtained from a sequence $\mathbb{D}\mathcal{X} \to \mathbb{D}\mathcal{O} \to y \to z$. Notice that, in the lower diagram, there is only one step that is entirely dynamic (the last step $y \to z$). This is the kind of single-step analysis covered by Theorem 3.

A comparison between the result of two single-step analyses, before and after the execution of a post-processing step (i.e., comparing the two diagrams from Figure 4a), corresponds to the question "Taking into account every possible path from the input that, going through system C, could lead to an output $y \in \mathcal{Y}$, is it safe to post-process y with S?" and the conclusion, given Theorem 3, is that it is safe *on average*. Here, even though $\mathrm{D} = \mathrm{B}\,;\mathrm{R}\,;\mathrm{S}$ involves a first post-processing step R, we are still considering every path through R. To illustrate that, recall Example 2a (pages 3 and 9): instead of focusing on a particular secret answer NO, a single-step analysis would consider every secret answer that could be mapped by the first post-processing step P to YES; if this was the kind of analysis we were interested on, then we could conclude that it is safe (on average) to replace P with $\mathrm{P}\,;\mathrm{S}$.

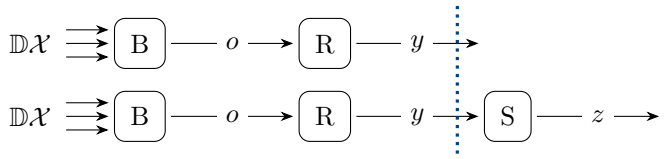
On the other hand, if we also focus on a concrete output $o \in \mathcal{O}$ from B, so that we are comparing the result of two analyses with semantics $\mathbb{D}\mathcal{X} \to o \to y$ vs. $\mathbb{D}\mathcal{X} \to o \to y \to z$, then the comparison now involves two dynamic steps, from $o \to y$ and then $y \to z$, which is *not* covered by Theorem 3. This is what we call a *multi-step* analysis, which Figure 4b depicts, and it is exactly the kind of analysis described in Example 2b, by focusing on a concrete original answer NO, instead of taking into account every possible original answer.

Hence, a comparison between two multi-step analyses can be interpreted as: "Taking into account every possible path from the input that, going through system B, could lead to an output $o \in \mathcal{O}$, which was then mapped to $y \in \mathcal{Y}$ by system R, is it safe to post-process y with S?". In this case the conclusion is that it is not necessarily safe, as Example 2b illustrates, and thus a stronger version of the data-processing inequality cannot be guaranteed in general. We reinforce, nonetheless, that the conclusions that we obtain from single- and multi-step analysis are not comparable, as their semantics are not compatible, and therefore one result does not invalidate the other.

A similar reasoning follows when assessing the information that leaks from a particular system, setting the post-processing channel $\mathrm{S} = \mathbb{1}$ (see Remark 3). In this case a single-step analysis corresponds to $\mathcal{L}^{\mathrm{st}}_{g}((\pi \triangleright \mathrm{C})_{X|y} \,\|\, \pi)$, as defined in Definition 10. This is the usual analysis we conduct, evaluating the threat to the secret before and after the adversary observes an execution of the system, from the adversary's point of view. This value is always non-negative, as stated in Theorem 2, and indicates that a rational adversary would always change their strategy once they observe an execution of the system.



(a) Single-step analysis.



(b) Multi-step analysis.

Fig. 4: The semantics of single- and multi-step analysis with respect to adversaries $\mathscr{A}^{\mathrm{C}}$ (upper diagram in each figure), who observes an output $y \in \mathcal{Y}$ from system $\mathrm{C} = \mathrm{B}\,;\mathrm{R}$, and $\mathscr{A}^{\mathrm{D}}$ (lower diagram), who observes $z \in \mathcal{Z}$ from system $\mathrm{D} = \mathrm{C}\,;\mathrm{S}$.

Nevertheless, a multi-step analysis is still meaningful *from the point of view of a data analyst* who has access to an intermediate result $o \in \mathcal{O}$ that is hidden to the adversary. Let channel $\mathrm{B} : \mathcal{X} \to \mathbb{D}\mathcal{O}$ model the system that produces the intermediate result, so that $(\pi \triangleright \mathrm{B})_{X|o}$ is the baseline. Then, we can define g- and ℓ-leakage more generally as, respectively,

$$V_g^{\mathrm{st}}((\pi \triangleright \mathrm{B})_{X|o} \,\|\, (\pi \triangleright \mathrm{C})_{X|y}) - V_g^{\mathrm{st}}((\pi \triangleright \mathrm{B})_{X|o} \,\|\, \pi) \qquad (16)$$

$$U_\ell^{\mathrm{st}}((\pi \triangleright \mathrm{B})_{X|o} \,\|\, \pi) - U_\ell^{\mathrm{st}}((\pi \triangleright \mathrm{B})_{X|o} \,\|\, (\pi \triangleright \mathrm{C})_{X|y}) \qquad (17)$$

This is *not* covered by Theorem 2, since here the baseline does not come from the channel that is available to the adversary (channel C). Therefore, it might be that, under this type of analysis, we observe negative information leakage. Such a result would mean that the adversary would perform better by constructing their strategy from their prior knowledge instead of using their posterior knowledge for that (at least, in this particular execution of the system). Yet, as already discussed, the adversary does not have access to the intermediate result $o \in \mathcal{O}$, and thus can only compute $\mathcal{L}_g^{\mathrm{st}}((\pi \triangleright \mathrm{C})_{X|y} \,\|\, \pi)$, which is always non-negative and would thus lead them to update their strategy based on their posterior. Hence, we argue that negative information leakage indicates a privacy improvement, as it corresponds to an "opportunity loss" for the adversary.

Bielova's Example 2 [17], using the Belief Tracking framework of Clarkson et al. [18], illustrates a case in which negative leakage can be observed. We note that Clarkson et al.'s framework occurs as a particular case of our formalisation. Clarkson et al. defines information leakage with respect to a particular secret input x^* as follows:

$$D_{KL}([x^*] \,\|\, \pi) - D_{KL}([x^*] \,\|\, (\pi \triangleright \mathrm{C})_{X|y}) \qquad (18)$$

But, recall from Theorem 1 that $D_{KL}(p \,\|\, q) = \mathcal{L}_{\ell_H}^{\mathrm{st}}(p \,\|\, q)$, and also notice that both terms in (18) involve subtracting $U_{\ell_H}^{\mathrm{st}}([x^*] \,\|\, [x^*]) = H([x^*]) = 0$. Therefore, we get that Clarkson et al.'s definition of information leakage is exactly our general definition given in (17), with $(\pi \triangleright \mathrm{B})_{X|o} = [x^*]$.

To further clarify the notion of a negative leakage, we now provide another example that illustrates this behaviour:

Example 5. Consider the scenario of a doctor (an "adversary") who is trying to diagnose a patient's disease (the "secret"). Suppose that, based on the symptoms described by the patient, the doctor knows that there are three possible diseases, but one of them is significantly more probable than the others. In this case the prior knowledge of the doctor can be represented, for instance, by the probability distribution $\pi = \{9/10, 1/20, 1/20\}$ ranging over diseases x_0, x_1, x_2. Based on this, the doctor would initially guess that the patient's disease is x_0. That is, $\mathrm{st}_{\mathbf{1}}^u(\pi) = [x_0]$, where the gain function is $\mathbf{1}$, as defined in (7).

Then, the doctor asks for a medical test to confirm their guess. The result of the test is either positive (P), which confirms that the patient has the disease x_0, or negative (N), which rules out x_0 as the possible disease, with 1% chance of error. The medical test can be represented by a channel $\mathrm{C} : \mathcal{X} \to \mathbb{D}\{\mathrm{P}, \mathrm{N}\}$. From that, and given the doctor's prior knowledge π, their posterior knowledge $[\pi \triangleright \mathrm{C}]$ would be

$$\begin{array}{c} \pi \\ x_0 \\ x_1 \\ x_2 \end{array} \begin{bmatrix} \frac{9}{10} \\ \frac{1}{20} \\ \frac{1}{20} \end{bmatrix} [\triangleright] \begin{array}{c} \mathrm{C} \\ x_0 \\ x_1 \\ x_2 \end{array} \begin{array}{c} \mathrm{P} \quad \mathrm{N} \\ \begin{bmatrix} \frac{99}{100} & \frac{1}{100} \\ \frac{1}{100} & \frac{99}{100} \\ \frac{1}{100} & \frac{99}{100} \end{bmatrix} \end{array} = \begin{array}{c} [\pi \triangleright \mathrm{C}] \\ x_0 \\ x_1 \\ x_2 \end{array} \begin{array}{c} \frac{223}{250}\,\mathrm{P} \quad \frac{27}{250}\,\mathrm{N} \\ \begin{bmatrix} \frac{891}{892} & \frac{1}{12} \\ \frac{1}{1784} & \frac{11}{24} \\ \frac{1}{1784} & \frac{11}{24} \end{bmatrix} \end{array}$$

If the output is N, the doctor's strategy a posteriori would be $\mathrm{st}_{\mathbf{1}}^u((\pi \triangleright \mathrm{C})_{X|\mathrm{N}}) = \{0, 1/2, 1/2\}$, since disease x_0 is now less likely than the others. Hence, the doctor would compute

$$\mathcal{L}_g^{\mathrm{st}}((\pi \triangleright \mathrm{C})_{X|\mathrm{N}} \,\|\, \pi) = 11/24 - 1/12 = 3/8$$

This indicates that the doctor should change their strategy to $\mathrm{st}_{\mathbf{1}}^u((\pi \triangleright \mathrm{C})_{X|\mathrm{N}})$. However, if we (as the "data analyst") knew that the disease is x_0 ("intermediate result"), which can be represented by a point distribution $[x_0]$, we would notice that

$$V_g^{\mathrm{st}}([x_0] \,\|\, (\pi \triangleright \mathrm{C})_{X|\mathrm{N}}) - V_g^{\mathrm{st}}([x_0] \,\|\, \pi) = 0 - 1 = -1 \qquad \square$$

Example 5 above shows that, under the dynamic scenario where the disease is x_0 and the result of the medical test was a false negative, the doctor's initial guess would be correct. But, because the doctor is a *rational* "adversary", the rational choice is to change their strategy once they observed the result of the medical test, and this choice leads them to an incorrect conclusion. This behaviour is captured by the negative leakage.

We reinforce that a negative leakage under a multi-step analysis is not an incorrect result. As shown in Example 5, it represents a rational adversary who, after observing an execution of the system, updated their prior strategy to a posterior strategy that, for a particular execution path, performs worse than the strategy they had a priori, and this behaviour is only observable through the lens of the data analyst.

VII. Related Work

The lack of a suitable definition of dynamic leakage is well known in the literature, and it is not particular to QIF. In [17] Bielova describes different measures of information leakage, and points out the need for a new measure that

evaluates information leakage with respect to one concrete output, regardless of the input. The approach that gets closer to such a measure is that of Belief Tracking [18], which quantifies the amount of information flow caused by changes in the accuracy of the adversary's belief, noting that the adversary's "belief" does not necessarily correspond to the "reality".

The drawback of Belief Tracking is that, in principle, it requires one to choose a reality against which we want to quantify information leakage (that is, to choose a secret input). Therefore, this does not fill the gap described in [17]. In contrast, our strategy-based formalisation proposed in Section IV does evaluate information leakage with respect to one concrete output without requiring a fixed secret input. Yet, Belief Tracking is similar to our proposal, in the sense that one has to choose a baseline, but it goes to the extreme of forcing this baseline to be a point distribution.

As shown in Section VI, Belief Tracking can be seen as a multi-step analysis, as its definition coincides with (17), which extends our formalisation of dynamic leakage to multiple dynamic steps. Although not explicitly defined in [18], it is possible to modify the Belief Tracking framework to single-step analysis — thus obtaining an instance of the information-leakage measure from Definition 10 — by taking the expectation over each possible baseline $[x]$, weighted by $(\pi \triangleright \mathrm{C})_{x|y}$.

Alvim et al. [28] have investigated the use of "strategies" in QIF, under "dynamic secrets". However similar, their terminology refers to secrets that change over time ("dynamic"), and in a scenario in which the adversary's prior is modelled by a hyper-distribution on secrets, that is, a distribution on "strategies" for generating secrets. Hence, there is little overlap between their contributions and those of this paper.

Privacy analyses such as those of Alvim et al. [29] and C. Soares et al. [7] can be viewed as a practical application of this dynamic perspective. They can be formalised following the framework presented in Section V for Example 4. More precisely, they follow a sort of mixed approach: they are dynamic with respect to the dataset observed by the attacker, but static in regard to the hints that the adversary could obtain.

VIII. Conclusion

In this paper we formalised a strategy-based definition of information leakage that solves the inconsistencies found when considering dynamic scenarios. More precisely, our definition preserves the properties of single-step monotonicity (i.e., by observing the output of a system, an adversary cannot lose information about the input) and of single-step data-processing inequality (i.e., one step of post-processing of the output of a system cannot create information; counter-intuitively, additional post-processing might "create" information). This approach builds on the idea raised in [18] and [28] (albeit with different purposes) that the adversary's belief does not necessarily correspond to the reality, and thus a definition of information leakage should take this distinction into account.

We proved that our formulation of strategy-based g-leakage preserves fundamental information-theoretic properties enjoyed by the static definition of leakage in the QIF framework. More precisely, we showed that the expectation of dynamic vulnerability corresponds to the traditional definition of expected vulnerability (equivalently, uncertainty). A similar property holds for the max-case. We also showed that our approach generalises that of Belief Tracking [18], as (i) it is parameterised by a gain (or loss) function modelling the attacker's goal; (ii) when instantiated with the loss function ℓ_H, it coincides with the Kullback-Leibler divergence, the measure adopted in Belief Tracking (Theorem 1); and (iii) it does not require one to choose a secret input as the reality.

We still require that a baseline is picked so that we can evaluate strategies against it, but the choice of a baseline is flexible, in the sense that one should choose the best baseline available (as long as it satisfies the knowledge ordering). In the cases where the goal is to quantify the information leakage of a system, the obvious choice is the adversary's posterior knowledge for one concrete output, whereas when comparing systems under a refinement relation, the baseline would be the least post-processed implementation available.

The fact that it is possible that the DPI does not hold in some dynamic scenarios (as illustrated in Example 2b) gives room to the possibility of a stronger notion of refinement (perhaps too strong to be satisfied in practice). We leave this for future work, as well as studying other interesting properties that have already been demonstrated for the static perspective, such as the behaviour of g-leakage measures under shifting and scaling of gain functions [15, Theorem 5.13].

Acknowledgment

Luigi D. C. Soares was supported by CNPq (Brazil), grant number 140359/2022-2, and by the International Cotutelle Macquarie University Research Excellence Scholarship.

References

[1] M. S. Alvim, N. Fernandes, A. McIver, and G. H. Nunes, "The privacy-utility trade-off in the topics api," in *Proceedings of the 2024 on ACM SIGSAC Conference on Computer and Communications Security*, 2024, pp. 1106–1120.

[2] M. S. Alvim, N. Fernandes, A. McIver, C. Morgan, and G. H. Nunes, "A novel analysis of utility in privacy pipelines, using Kronecker products and quantitative information flow," in *Proc. SIGSAC*, 2023.

[3] M. Jurado, C. Palamidessi, and G. Smith, "A formal information-theoretic leakage analysis of order-revealing encryption," in *Proc. CSF*, 2021.

[4] N. Fernandes, M. Dras, and A. McIver, "Processing text for privacy: an information flow perspective," in *Proc. FM*, 2018.

[5] M. S. Alvim, M. E. Andrés, K. Chatzikokolakis, P. Degano, and C. Palamidessi, "On the information leakage of differentially-private mechanisms," *Journal Computer Security*, vol. 23, no. 4, pp. 427–469, 2015. [Online]. Available: https://doi.org/10.3233/JCS-150528

[6] K. Chatzikokolakis, N. Fernandes, and C. Palamidessi, "Comparing systems: Max-case refinement orders and application to differential privacy," in *Proc. CSF*, 2019.

[7] L. D. C. Soares, M. S. Alvim, D. Bu, N. Fernandes, and Y. Liao, "Formal Privacy Analyses for Open Banking," in *Formal Methods: Foundations and Applications*, S. C. Nogueira and C. Teodorov, Eds. Cham: Springer Nature Switzerland, 2025, vol. 15403, pp. 171–193, series Title: Lecture Notes in Computer Science. [Online]. Available: https://link.springer.com/10.1007/978-3-031-78116-2_11

[8] R. M. Silva, G. C. M. Gomes, M. S. Alvim, and M. A. Gonçalves, "How to build high quality L2R training data: Unsupervised compression-based selective sampling for learning to rank," *Information Sciences*, vol. 601, 2022.

[9] F. Viegas, M. S. Alvim, S. D. Canuto, T. Rosa, M. A. Gonçalves, and L. Rocha, "Exploiting semantic relationships for unsupervised expansion of sentiment lexicons," *Information Systems*, vol. 94, 2020.

[10] J. M. Pimentel, M. S. Alvim, M. F. M. Campos, and D. G. Macharet, "Information-driven rapidly-exploring random tree for efficient environment exploration," *J. Intell. Robotic Syst.*, vol. 91, no. 2, 2018.

[11] F. Moraes, M. S. Alvim, and R. L. T. Santos, "Modeling information flow in dynamic information retrieval," in *Proc. ICTR*, J. Kamps, E. Kanoulas, M. de Rijke, H. Fang, and E. Yilmaz, Eds., 2017.

[12] M. S. Alvim, N. Fernandes, B. D. Nogueira, C. Palamidessi, and T. V. A. Silva, "On the Duality of Privacy and Fairness (Extended Abstract)," in *Proc. CADE*, 2023.

[13] S. A. Mário, K. Chatzikokolakis, C. Palamidessi, and G. Smith, "Measuring information leakage using generalized gain functions," in *2012 IEEE 25th Computer Security Foundations Symposium*. IEEE, 2012, pp. 265–279.

[14] M. S. Alvim, K. Chatzikokolakis, A. McIver, C. Morgan, C. Palamidessi, and G. Smith, "An axiomatization of information flow measures," *Theoretical Computer Science*, vol. 777, pp. 32–54, 2019.

[15] ——, *The Science of Quantitative Information Flow*. Springer, 2020.

[16] T. M. Cover, *Elements of information theory*. John Wiley & Sons, 1999.

[17] N. Bielova, "Short Paper: Dynamic leakage: A Need for a New Quantitative Information Flow Measure," in *Proceedings of the 2016 ACM Workshop on Programming Languages and Analysis for Security*. Vienna Austria: ACM, Oct. 2016, pp. 83–88. [Online]. Available: https://dl.acm.org/doi/10.1145/2993600.2993607

[18] M. R. Clarkson, A. C. Myers, and F. B. Schneider, "Quantifying information flow with beliefs," *Journal of Computer Security*, vol. 17, no. 5, pp. 655–701, Oct. 2009. [Online]. Available: https://www.medra.org/servlet/aliasResolver?alias=iospress&doi=10.3233/JCS-2009-0353

[19] M. S. Alvim, K. Chatzikokolakis, C. Palamidessi, and G. Smith, "Measuring Information Leakage Using Generalized Gain Functions," in *2012 IEEE 25th Computer Security Foundations Symposium*. Cambridge, MA, USA: IEEE, Jun. 2012, pp. 265–279. [Online]. Available: http://ieeexplore.ieee.org/document/6266165/

[20] M. Thomas and A. T. Joy, *Elements of information theory*. Wiley-Interscience, 2006.

[21] G. Smith, "On the Foundations of Quantitative Information Flow," in *Foundations of Software Science and Computational Structures*, L. De Alfaro, Ed. Berlin, Heidelberg: Springer Berlin Heidelberg, 2009, vol. 5504, pp. 288–302, series Title: Lecture Notes in Computer Science. [Online]. Available: http://link.springer.com/10.1007/978-3-642-00596-1_21

[22] K. Chatzikokolakis, N. Fernandes, and C. Palamidessi, "Comparing Systems: Max-Case Refinement Orders and Application to Differential Privacy," in *2019 IEEE 32nd Computer Security Foundations Symposium (CSF)*. Hoboken, NJ, USA: IEEE, Jun. 2019, pp. 442–44 215. [Online]. Available: https://ieeexplore.ieee.org/document/8823711/

[23] T. Dalenius, "Finding a needle in a haystack or identifying anonymous census records," *Journal of official statistics*, vol. 2, no. 3, p. 329, 1986.

[24] P. Samarati and L. Sweeney, "Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression," 1998.

[25] L. Sweeney, "Simple demographics often identify people uniquely," *Health (San Francisco)*, vol. 671, no. 2000, pp. 1–34, 2000.

[26] A. Narayanan and V. Shmatikov, "Robust De-anonymization of Large Sparse Datasets," in *Proc. of S&P*, 2008, pp. 111–125.

[27] A. Américo, M. S. Alvim, and A. McIver, "An algebraic approach for reasoning about information flow," in *Formal Methods*, K. Havelund, J. Peleska, B. Roscoe, and E. de Vink, Eds. Cham: Springer International Publishing, 2018, pp. 55–72.

[28] M. S. Alvim, P. Mardziel, and M. W. Hicks, "Quantifying vulnerability of secret generation using hyper-distributions," in *Principles of Security and Trust - 6th International Conference, POST 2017, Held as Part of the European Joint Conferences on Theory and Practice of Software, ETAPS 2017, Uppsala, Sweden, April 22-29, 2017, Proceedings*, ser. Lecture Notes in Computer Science, M. Maffei and M. Ryan, Eds., vol. 10204. Springer, 2017, pp. 26–48. [Online]. Available: https://doi.org/10.1007/978-3-662-54455-6_2

[29] M. S. Alvim, N. Fernandes, A. McIver, C. Morgan, and G. H. Nunes, "Flexible and scalable privacy assessment for very large datasets, with an application to official governmental microdata," *Proc. Priv. Enhancing Technol.*, vol. 2022, no. 4, pp. 378–399, 2022. [Online]. Available: https://doi.org/10.56553/popets-2022-0114

APPENDIX A
PROOFS OMITTED FROM THE MAIN BODY

A. Proofs omitted from Section IV-B

In Section IV we claimed that the expected vulnerability caused by an adversary who selects one of the possible optimal strategies at random converges to the vulnerability caused by an adversary who picks the uniform strategy. This was illustrated in Example 3, by limiting the strategies available to the adversary to those that only assign to the optimal actions probabilities with exactly $n = 1$ decimal place. In this case there were 66 optimal strategies available to the adversary, and the average over the vulnerability caused by each strategy, as defined in (12), was $1/3$, exactly the same as the vulnerability caused by the uniform strategy. Then, we affirmed that this phenomenon is actually observed for any $n \geq 0$ that we choose as the number of decimal places allowed in the definition of strategies. We now formalise and prove this statement:

Theorem A.1. *Let $q : \mathbb{D}\mathcal{X}$ be the adversary's belief about the secret and $g : \mathcal{W} \times \mathcal{X} \to \mathbb{R}$ be the gain function modelling the adversary's goal. Furthermore, let $\mathrm{st}_g^n(q)$, $n \geq 0$, be the (finite) set of optimal strategies available to the adversary, such that the probability assigned by any strategy in $\mathrm{st}_g^n(q)$ to any action $w \in \mathcal{W}$ is restricted to exactly n decimal places. Finally, let $\mathrm{st}_g^u(q)$ be the uniform strategy. Then, assuming that the set of optimal actions with respect to q is finite,*

$$\lim_{n\to\infty} \sum_{s\in \mathrm{st}_g^n(q)} \frac{1}{|\mathrm{st}_g^n(q)|} \sum_{w\in\mathcal{W}} s_w \sum_{x\in\mathcal{X}} p_x\, g(w,x) \qquad (19)$$
$$= \sum_{w\in\mathcal{W}} \mathrm{st}_g^u(q)_w \sum_{x\in\mathcal{X}} p_x\, g(w,x)$$

Dually, for any loss function $\ell : \mathcal{W} \times \mathcal{X} \to \mathbb{R}_{\geq 0} \cup \{\infty\}$,

$$\lim_{n\to\infty} \sum_{s\in \mathrm{st}_\ell^n(q)} \frac{1}{|\mathrm{st}_\ell^n(q)|} \sum_{w\in\mathcal{W}} s_w \sum_{x\in\mathcal{X}} p_x\, \ell(w,x) \qquad (20)$$
$$= \sum_{w\in\mathcal{W}} \mathrm{st}_\ell^u(q)_w \sum_{x\in\mathcal{X}} p_x\, \ell(w,x)$$

Proof. Recall from Definition 8 that an (optimal) adversarial strategy $s \in \mathbb{D}\mathcal{W}$, with respect to a state of knowledge $q : \mathbb{D}\mathcal{X}$ (the adversary's belief) and a gain function $g : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$, is a probability distribution on the set of actions $\mathcal{W}$, such that

$$s_{w^*} > 0 \implies w^* \in \operatorname*{argmax}_{w\in\mathcal{W}} \sum_{x\in\mathcal{X}} q_x\, g(w,x) \qquad (21)$$

and that $\mathrm{st}_g(q)$ is the set of all strategies satisfying the above.

Let $\mathcal{W}^* = \operatorname{argmax}_{w\in\mathcal{W}} \sum_{x\in\mathcal{X}} q_x\, g(w,x)$ be the set of optimal actions (finite, as per hypothesis) with respect to the state of knowledge q and gain function g, such that the uniform strategy over all optimal actions is defined as follows:

$$\mathrm{st}_g^u(q)_w \stackrel{\text{def}}{=} \frac{1}{|\mathcal{W}^*|} \text{ if } w \in \mathcal{W}^* \text{ else } 0 \qquad (22)$$

Fix an arbitrary $n \geq 0$ for the number of decimal places allowed in the definition of strategies. Then, consider the following matrix, where each column is one of the (finitely many) strategies $s^j \in \mathrm{st}_g^n(q)$ and each row is labelled by one of the (finitely many) optimal actions $w_i^* \in \mathcal{W}^*$:

$$\begin{array}{c} \\ w_0^* \\ w_1^* \\ \vdots \\ w_{|\mathcal{W}^*|-1}^* \end{array} \begin{array}{c} \begin{array}{cccc} s^0 & s^1 & \cdots & s^{|\mathrm{st}_g^n(q)|-1} \end{array} \\ \begin{bmatrix} s^0_{w_0^*} & s^1_{w_0^*} & \cdots & s^{|\mathrm{st}_g^n(q)|-1}_{w_0^*} \\ s^0_{w_1^*} & s^1_{w_1^*} & \cdots & s^{|\mathrm{st}_g^n(q)|-1}_{w_1^*} \\ \vdots & \vdots & \vdots & \vdots \\ s^0_{w^*_{|\mathcal{W}^*|-1}} & s^1_{w^*_{|\mathcal{W}^*|-1}} & \cdots & s^{|\mathrm{st}_g^n(q)|-1}_{w^*_{|\mathcal{W}^*|-1}} \end{bmatrix} \end{array}$$

Notice that, in the matrix above, the two following facts must hold: (i) each column must add up to 1, since it corresponds to a probability distribution on optimal actions, so the sum of all cells in the matrix must equal the number $|\mathrm{st}_g^n(q)|$ of optimal strategies; and (ii) each row contains *all* probabilities that could be assigned by a strategy to the corresponding action and therefore, by symmetry, the sum of each row must be the same, i.e., it must be the sum $|\mathrm{st}_g^n(q)|$ of all cells in the matrix divided by the number $|\mathcal{W}^*|$ of rows. Putting these two facts together, we conclude that the sum of each row i must be:

$$\sum_{j=0}^{|\mathrm{st}_g^n(q)|-1} s^j_{w_i^*} = \frac{|\mathrm{st}_g^n(q)|}{|\mathcal{W}^*|} \qquad (23)$$

Equipped with (23), we can then reason as follows:

$$\begin{aligned}
&\lim_{n\to\infty} \sum_{s\in \mathrm{st}_g^n(q)} \frac{1}{|\mathrm{st}_g^n(q)|} \sum_{w\in\mathcal{W}} s_w \sum_{x\in\mathcal{X}} p_x\, g(w,x) \\
&= \lim_{n\to\infty} \sum_{w^*\in\mathcal{W}^*} \sum_{x\in\mathcal{X}} p_x\, g(w^*,x) \\
&\quad \left(\frac{1}{|\mathrm{st}_g^n(q)|} \sum_{j=0}^{|\mathrm{st}_g^n(q)|-1} s^j_{w^*} \right) && \text{\# Rearrange terms; \# Rewrite indices} \\
&= \lim_{n\to\infty} \sum_{w^*\in\mathcal{W}^*} \sum_{x\in\mathcal{X}} p_x\, g(w^*,x) \left(\frac{|\mathrm{st}_g^n(q)|}{|\mathrm{st}_g^n(q)|\,|\mathcal{W}^*|} \right) && \text{\# See (23)} \\
&= \lim_{n\to\infty} \sum_{w^*\in\mathcal{W}^*} \sum_{x\in\mathcal{X}} p_x\, g(w^*,x) \frac{1}{|\mathcal{W}^*|} \\
&= \sum_{w^*\in\mathcal{W}^*} \frac{1}{|\mathcal{W}^*|} \sum_{x\in\mathcal{X}} p_x\, g(w^*,x) && \text{\# Rearrange terms; \# No depedency on } n \\
&= \sum_{w\in\mathcal{W}} \mathrm{st}_g^u(q)_w \sum_{x\in\mathcal{X}} p_x\, g(w^*,x) && \text{\# See (22)}
\end{aligned}$$

□

B. Proofs omitted from Section IV-C

Lemma 1 can be seen as a particular case of a more general lemma, stated as follows:

Lemma A.1. *Consider two systems modelled by channels $\mathrm{C} : \mathcal{X} \to \mathbb{D}\mathcal{Y}$ and $\mathrm{D} : \mathcal{X} \to \mathbb{D}\mathcal{Z}$ such that $\mathrm{C} \sqsubseteq \mathrm{D}$, and recall that $\mathrm{D} = \mathrm{C}\,;\mathrm{R}$. It follows that, for any prior knowledge $\pi : \mathbb{D}\mathcal{X}$, and outputs $y \in \mathcal{Y}$ and $z \in \mathcal{Z}$ such that $\mathrm{R}_{y,z} > 0$, $U_{\ell_H}^{\mathrm{st}}((\pi \triangleright \mathrm{C})_{X|y} \,\|(\pi \triangleright \mathrm{D})_{X|z})$ is finite and non-negative.*

Proof. To show that, we just need to demonstrate that if $(\pi \triangleright \mathrm{D})_{x|z} = 0$, then $(\pi \triangleright \mathrm{C})_{x|y} = 0$ as well, so that whenever $\ell_H((\pi \triangleright \mathrm{D})_{x|z}, x) = \infty$, $(\pi \triangleright \mathrm{C})_{x|y} = 0$ and thus $(\pi \triangleright \mathrm{C})_{x|y}\, \ell_H((\pi \triangleright \mathrm{D})_{x|z}, x) = 0$.

If $(\pi \triangleright \mathrm{D})_{x|z} = 0$, we have (from Defn 7, Eqn 6 and Eqn 4)

$$(\pi \triangleright \mathrm{D})_{x|z} = \frac{\pi_x\, (\mathrm{C}\,;\mathrm{R})_{x,z}}{(\pi \triangleright \mathrm{C}\,;\mathrm{R})_z} = \frac{\pi_x \sum_{y'} \mathrm{C}_{x,y'}\, \mathrm{R}_{y',z}}{(\pi \triangleright \mathrm{C}\,;\mathrm{R})_z} = 0$$

and we know that $\mathrm{R}_{y,z} > 0$. Therefore, either $\pi_x = 0$ or $\mathrm{C}_{x,y} = 0$, and both cases imply that $(\pi \triangleright \mathrm{C})_{x|z} = 0$. □

Lemma 1. *For any prior knowledge $\pi : \mathbb{D}\mathcal{X}$ and channel $\mathrm{C} : \mathcal{X} \to \mathbb{D}\mathcal{Y}$, it follows that both the strategy-based dynamic prior ℓ_H-uncertainty $U^{\mathrm{st}}_{\ell_H}((\pi \triangleright \mathrm{C})_{X|y} \,\|\, \pi)$ and the strategy-based dynamic posterior ℓ_H-uncertainty $U^{\mathrm{st}}_{\ell_H}((\pi \triangleright \mathrm{C})_{X|y} \,\|\, (\pi \triangleright \mathrm{C})_{X|y})$ are finite and non-negative.*

Proof. This follows directly from Lemma A.1, setting $\mathrm{R} = \mathbb{1}$ so that $(\pi \triangleright \mathrm{D})_{X|z} = (\pi \triangleright \mathbb{1})_{X|\perp} = \pi$. □

Theorem 1. *Let $q : \mathbb{D}\mathcal{X}$ be the adversary's belief and $p : \mathbb{D}\mathcal{X}$ be the baseline. Then, $\mathcal{L}^{\mathrm{st}}_{\ell_H}(p \,\|\, q) = D_{KL}(p \,\|\, q)$.*

Proof. Using Gibbs' inequality, we can show that there is only one optimal action that minimises the expected loss with respect to the loss function ℓ_H, as defined in (8). That is, for any states of knowledge $\pi : \mathbb{D}\mathcal{X}$ and $\delta : \mathbb{D}\mathcal{X}$, it follows that

$$\sum_{x \in \mathcal{X}} \delta_x\, \ell_H(\delta, x) < \sum_{x \in \mathcal{X}} \delta_x\, \ell_H(\pi, x) \quad \forall \pi \neq \delta \tag{24}$$

With this, we can reason as follows:

$$\begin{aligned}
&\mathcal{L}^{\mathrm{st}}_{\ell_H}(p \,\|\, q) \\
&= U^{\mathrm{st}}_{\ell_H}(p \,\|\, q) - U_{\ell_H}(p) && \text{\# Definition 10} \\
&= \left(\sum_{w \in \mathcal{W}} \mathrm{st}^u_{\ell_H}(q)_w \sum_{x \in \mathcal{X}} p_x\, \ell_H(w, x)\right) - U_{\ell_H}(p) && \text{\# Definition 9} \\
&= \left(\sum_{x \in \mathcal{X}} p_x\, \ell_H(q, x)\right) - U_{\ell_H}(p) && \text{\# From (24) and Definition 8,} \\
& && \text{\# } \mathrm{st}^u_{\ell_H}(q) = [q], \\
& && \text{\# with [q] as defined in (13)} \\
&= \left(\sum_{x \in \mathcal{X}} p_x\, \ell_H(q, x)\right) - \left(\min_{w \in \mathcal{W}} \sum_{x \in \mathcal{X}} p_x\, \ell_H(w, x)\right) && \text{\# See (10)} \\
&= \left(\sum_{x \in \mathcal{X}} p_x\, \ell_H(q, x)\right) - \left(\sum_{x \in \mathcal{X}} p_x\, \ell_H(p, x)\right) && \text{\# From (24),} \\
& && \text{\# } \operatorname{argmin}_w = p \\
&= \sum_{x \in \mathcal{X}} p_x\, (\ell_H(q, x) - \ell_H(p, x)) && \text{\# Merge sum} \\
&= \sum_{x \in \mathcal{X}} p_x\, (-\log_2 q_x + \log_2 p_x) && \text{\# From (8)} \\
&= \sum_{x \in \mathcal{X}} p_x \log_2 \frac{p_x}{q_x} && \text{\# Log difference} \\
&= D_{KL}(p \,\|\, q) && \text{\# From (14)}
\end{aligned}$$

□

C. Proofs omitted from Section IV-E

Lemma 2. *Let $p : \mathbb{D}\mathcal{X}$ be both the adversary's belief and the baseline. Then, for any gain function $g : \mathcal{W} \times \mathcal{X} \to \mathbb{R}$,*

$$V^{\mathrm{st}}_g(p \,\|\, p) = V_g(p)$$

Dually, for any loss function $\ell : \mathcal{W} \times \mathcal{X} \to \mathbb{R}_{\geq 0} \cup \{\infty\}$,

$$U^{\mathrm{st}}_\ell(p \,\|\, p) = U_\ell(p)$$

Proof. Let $\mathrm{supp}(p)$ be the support of a probability distribution p. Then, for any state of knowledge $p : \mathbb{D}\mathcal{X}$, any strategy $s \in \mathrm{st}_g(p)$ and any gain function $g : \mathcal{W} \times \mathcal{X} \to \mathbb{R}$, we have

$$\begin{aligned}
&\sum_{w \in \mathcal{W}} s_w \sum_{x \in \mathcal{X}} p_x\, g(w, x) && \text{\# From (12)} \\
&= \sum_{w^* \in \mathrm{supp}(s)} s_{w^*} \sum_{x \in \mathcal{X}} p_x\, g(w^*, x) \\
&= \sum_{w^* \in \mathrm{supp}(s)} s_{w^*} \max_{w \in \mathcal{W}} \sum_{x \in \mathcal{X}} p_x\, g(w, x) && (\star) \\
&= \sum_{w^* \in \mathrm{supp}(s)} s_{w^*}\, V_g(p) && \text{\# From (9)} \\
&= V_g(p) \sum_{w^* \in \mathrm{supp}(s)} s_{w^*} && \text{\# Rearrange terms} \\
&= V_g(p) && \text{\# } s \text{ is a probility distribution;} \\
& && \text{\# } \forall w \notin \mathrm{supp}(s) \colon s_w = 0
\end{aligned}$$

noting that in $(\star)$ every action $w^* \in \mathrm{supp}(s)$ maximises $\sum_{x \in \mathcal{X}} p_x\, g(w^*, x)$. The above holds for any (optimal) strategy $s \in \mathrm{st}_g(p)$, which includes the uniform strategy adopted in Defintion 9. The proof for the second part (uncertainty) can be obtained in the same way, replacing $\max$ with $\min$. □

Lemma 3. *Let $q : \mathbb{D}\mathcal{X}$ be the adversary's belief and $p : \mathbb{D}\mathcal{X}$ be the baseline. Then, for any gain function $g : \mathcal{W} \times \mathcal{X} \to \mathbb{R}$,*

$$V^{\mathrm{st}}_g(p \,\|\, q) \leq V^{\mathrm{st}}_g(p \,\|\, p) = V_g(p)$$

with equality holding when $q = p$. Dually, for any loss function $\ell : \mathcal{W} \times \mathcal{X} \to \mathbb{R}_{\geq 0} \cup \{\infty\}$, it follows that

$$U^{\mathrm{st}}_\ell(p \,\|\, q) \geq U^{\mathrm{st}}_\ell(p \,\|\, p) = U_\ell(p)$$

Proof. The equality $V^{\mathrm{st}}_g(p \,\|\, p) = V_g(p)$ (right-hand side) follows from Lemma 2. Then, for any $s \in \mathrm{st}_g(q)$, we have

$$\begin{aligned}
&\sum_{w \in \mathcal{W}} s_w \sum_{x \in \mathcal{X}} p_x\, g(w, x) && \text{\# From (12)} \\
&\leq \sum_{w \in \mathcal{W}} s_w \max_{w' \in \mathcal{W}} \sum_{x \in \mathcal{X}} p_x\, g(w', x) \\
&= \sum_{w \in \mathcal{W}} s_w\, V_g(p) && \text{\# From (9)} \\
&= V_g(p) \sum_{w \in \mathcal{W}} s_w && \text{\# Rearrange terms} \\
&= V_g(p) && \text{\# } s \text{ is a probability distribution}
\end{aligned}$$

Since the above holds for any strategy $s \in \mathrm{st}_g(q)$, it also holds for the uniform strategy $\mathrm{st}^u_g(q)$ adopted in Definition 9. The proof for the second part (ℓ-uncertainty) can be obtained in the same way, replacing $\max$ with $\min$ and $\leq$ with $\geq$. □

Theorem 2 (Single-step monotonicity)**.** *Let $\pi : \mathbb{D}\mathcal{X}$ be the adversary's prior belief, and consider a channel $\mathrm{C} : \mathcal{X} \to \mathbb{D}\mathcal{Y}$ producing $y \in \mathcal{Y}$. Then, for any gain function $g : \mathcal{W} \times \mathcal{X} \to \mathbb{R}$,*

$$V_g^{\mathrm{st}}((\pi \triangleright \mathrm{C})_{X|y} \,\|\, \pi) \leq V_g^{\mathrm{st}}((\pi \triangleright \mathrm{C})_{X|y} \,\|\, (\pi \triangleright \mathrm{C})_{X|y})$$

Dually, for any loss function $\ell : \mathcal{W} \times \mathcal{X} \to \mathbb{R}_{\geq 0} \cup \{\infty\}$,

$$U_\ell^{\mathrm{st}}((\pi \triangleright \mathrm{C})_{X|y} \,\|\, \pi) \geq U_\ell^{\mathrm{st}}((\pi \triangleright \mathrm{C})_{X|y} \,\|\, (\pi \triangleright \mathrm{C})_{X|y})$$

Proof. This follows directly from Lemma 3, by setting the baseline p as $(\pi \triangleright \mathrm{C})_{X|y}$ and the adversary's belief q as π. □

Theorem 3 (Single-step DPI)**.** *Let $\pi : \mathbb{D}\mathcal{X}$ be the adversary's prior belief, and consider two channels $\mathrm{C} : \mathcal{X} \to \mathbb{D}\mathcal{Y}$ and $\mathrm{D} : \mathcal{X} \to \mathbb{D}\mathcal{Z}$ such that $\mathrm{C} \sqsubseteq \mathrm{D}$. Let $y \in \mathcal{Y}$ and $z \in \mathcal{Z}$ such that $\mathrm{R}_{y,z} > 0$. Then, for any gain function $g : \mathcal{W} \times \mathcal{X} \to \mathbb{R}$,*

$$V_g^{\mathrm{st}}((\pi \triangleright \mathrm{C})_{X|y} \,\|\, (\pi \triangleright \mathrm{D})_{X|z}) \leq V_g^{\mathrm{st}}((\pi \triangleright \mathrm{C})_{X|y} \,\|\, (\pi \triangleright \mathrm{C})_{X|y})$$

Dually, for any loss function $\ell : \mathcal{W} \times \mathcal{X} \to \mathbb{R}_{\geq 0} \cup \{\infty\}$,

$$U_\ell^{\mathrm{st}}((\pi \triangleright \mathrm{C})_{X|y} \,\|\, (\pi \triangleright \mathrm{D})_{X|z}) \geq U_\ell^{\mathrm{st}}((\pi \triangleright \mathrm{C})_{X|y} \,\|\, (\pi \triangleright \mathrm{C})_{X|y})$$

Proof. This follows directly from Lemma 3, by setting the baseline p as $(\pi \triangleright \mathrm{C})_{X|y}$ and the belief q as $(\pi \triangleright \mathrm{D})_{X|z}$. □

D. Proofs omitted from Section IV-F

Theorem 5. *Let $\pi : \mathbb{D}\mathcal{X}$ be the adversary's prior belief, and consider a channel $\mathrm{C} : \mathcal{X} \to \mathbb{D}\mathcal{Y}$ modelling the system of interest. Then, for any gain function $g : \mathcal{W} \times \mathcal{X} \to \mathbb{R}$,*

$$\sum_{y \in \mathcal{Y}} (\pi \triangleright \mathrm{C})_y \, V_g^{\mathrm{st}}((\pi \triangleright \mathrm{C})_{X|y} \,\|\, (\pi \triangleright \mathrm{C})_{X|y}) = V_g[\pi \triangleright \mathrm{C}]$$

Dually, for any loss function $\ell : \mathcal{W} \times \mathcal{X} \to \mathbb{R}_{\geq 0} \cup \{\infty\}$,

$$\sum_{y \in \mathcal{Y}} (\pi \triangleright \mathrm{C})_y \, U_\ell^{\mathrm{st}}((\pi \triangleright \mathrm{C})_{X|y} \,\|\, (\pi \triangleright \mathrm{C})_{X|y}) = U_\ell[\pi \triangleright \mathrm{C}]$$

Proof. This follows immediately from Lemma 2. □

Theorem 6. *Let $\pi : \mathbb{D}\mathcal{X}$ be the adversary's prior belief, and consider a channel $\mathrm{C} : \mathcal{X} \to \mathbb{D}\mathcal{Y}$ modelling the system of interest. Then, for any gain function $g : \mathcal{W} \times \mathcal{X} \to \mathbb{R}$,*

$$\max_{y \in \mathcal{Y}} V_g^{\mathrm{st}}((\pi \triangleright \mathrm{C})_{X|y} \,\|\, (\pi \triangleright \mathrm{C})_{X|y}) = V_g^{\max}[\pi \triangleright \mathrm{C}]$$

Dually, for any loss function $\ell : \mathcal{W} \times \mathcal{X} \to \mathbb{R}_{\geq 0} \cup \{\infty\}$,

$$\min_{y \in \mathcal{Y}} U_\ell^{\mathrm{st}}((\pi \triangleright \mathrm{C})_{X|y} \,\|\, (\pi \triangleright \mathrm{C})_{X|y}) = U_\ell^{\min}[\pi \triangleright \mathrm{C}]$$

Proof. This follows immediately from Lemma 2. □

Theorem 7. *Let $\pi : \mathbb{D}\mathcal{X}$ be the adversary's prior belief, and consider a channel $\mathrm{C} : \mathcal{X} \to \mathbb{D}\mathcal{Y}$ modelling the system of interest. Then, for any gain function $g : \mathcal{W} \times \mathcal{X} \to \mathbb{R}$,*

$$\sum_{y \in \mathcal{Y}} (\pi \triangleright \mathrm{C})_y \, V_g^{\mathrm{st}}((\pi \triangleright \mathrm{C})_{X|y} \,\|\, \pi) = V_g(\pi)$$

Dually, for any loss function $\ell : \mathcal{W} \times \mathcal{X} \to \mathbb{R}_{\geq 0} \cup \{\infty\}$,

$$\sum_{y \in \mathcal{Y}} (\pi \triangleright \mathrm{C})_y \, U_\ell^{\mathrm{st}}((\pi \triangleright \mathrm{C})_{X|y} \,\|\, \pi) = U_\ell(\pi)$$

Proof. For any observation $y \in \mathcal{Y}$, any gain function $g : \mathcal{W} \times \mathcal{X} \to \mathbb{R}$ and any strategy $s \in \mathrm{st}_g(\pi)$, it follows that

$$\begin{aligned}
&\sum_{y \in \mathcal{Y}} (\pi \triangleright \mathrm{C})_y \sum_{w \in \mathcal{W}} s_w \sum_{x \in \mathcal{X}} (\pi \triangleright \mathrm{C})_{x|y} \, g(w,x) && \text{\# From (12)} \\
= &\sum_{w \in \mathcal{W}} s_w \\
&\sum_{y \in \mathcal{Y}} \sum_{x \in \mathcal{X}} (\pi \triangleright \mathrm{C})_y \, (\pi \triangleright \mathrm{C})_{x|y} \, g(w,x) && \text{\# Rearrange terms} \\
= &\sum_{w \in \mathcal{W}} s_w \\
&\sum_{y \in \mathcal{Y}} \sum_{x \in \mathcal{X}} (\pi \triangleright \mathrm{C})_{x,y} \, g(w,x) && \text{\# From posterior (6) to joint} \\
= &\sum_{w \in \mathcal{W}} s_w \\
&\sum_{y \in \mathcal{Y}} \sum_{x \in \mathcal{X}} \pi_x \, \mathrm{C}_{x,y} \, g(w,x) && \text{\# From joint (4) to prior \& channel} \\
= &\sum_{w \in \mathcal{W}} s_w \sum_{x \in \mathcal{X}} \pi_x \, g(w,x) \sum_{y \in \mathcal{Y}} \mathrm{C}_{x,y} && \text{\# Rearrange terms} \\
= &\sum_{w \in \mathcal{W}} s_w \sum_{x \in \mathcal{X}} \pi_x \, g(w,x) && \text{\# Row } x \text{ is a probability distribution} \\
= &\; V_g(\pi) && \text{\# Proof of Lemma 2}
\end{aligned}$$

Since the above holds for any strategy $s \in \mathrm{st}_g(\pi)$, it also holds for the uniform strategy $\mathrm{st}_g^u(\pi)$ adopted in Definition 9. The proof for the second part (ℓ-uncertainty) can be obtained in a similar way, and thus is omitted. □

To prove Theorems 8 and 9, we first show the "squashing" effect for a fixed output $z \in \mathcal{Z}$ from channel D:

Lemma A.2. *Let $\pi : \mathbb{D}\mathcal{X}$ be the adversary's prior belief, and consider two channels $\mathrm{C} : \mathcal{X} \to \mathbb{D}\mathcal{Y}$ and $\mathrm{D} : \mathcal{X} \to \mathbb{D}\mathcal{Z}$, $\mathrm{C} \sqsubseteq \mathrm{D}$, so that $\mathrm{D} = \mathrm{C} \,;\, \mathrm{R}$. Moreover, let $z \in \mathcal{Z}$ be an output from* D*. For any gain function $g : \mathcal{W} \times \mathcal{X} \to \mathbb{R}$, the* dynamic *posterior g-vulnerability $V_g((\pi \triangleright \mathrm{D})_{X|z})$ is recovered as*

$$\frac{1}{(\pi \triangleright \mathrm{D})_z} \sum_{y \in \mathcal{Y}} (\pi \triangleright \mathrm{C})_y \, \mathrm{R}_{y,z} \, V_g^{\mathrm{st}}((\pi \triangleright \mathrm{C})_{X|y} \,\|\, (\pi \triangleright \mathrm{D})_{X|z})$$

Dually, for any loss function $\ell : \mathcal{W} \times \mathcal{X} \to \mathbb{R}_{\geq 0} \cup \{\infty\}$, the dynamic *posterior ℓ-uncertainty $U_\ell((\pi \triangleright \mathrm{D})_{X|z})$ equals to*

$$\frac{1}{(\pi \triangleright \mathrm{D})_z} \sum_{y \in \mathcal{Y}} (\pi \triangleright \mathrm{C})_y \, \mathrm{R}_{y,z} \, U_\ell^{\mathrm{st}}((\pi \triangleright \mathrm{C})_{X|y} \,\|\, (\pi \triangleright \mathrm{D})_{X|z})$$

Proof. For any observation $z \in \mathcal{Z}$, any gain function $g : \mathcal{W} \times \mathcal{X} \to \mathbb{R}$ and any strategy $s \in \mathrm{st}_g((\pi \triangleright \mathrm{D})_{X|z})$, it follows that

$$\begin{aligned}
&\frac{1}{(\pi \triangleright \mathrm{D})_z} \sum_{y \in \mathcal{Y}} (\pi \triangleright \mathrm{C})_y \, \mathrm{R}_{y,z} \\
&\sum_{w \in \mathcal{W}} s_w \sum_{x \in \mathcal{X}} (\pi \triangleright \mathrm{C})_{x|y} \, g(w,x) && \text{\# From (12)} \\
= &\frac{1}{(\pi \triangleright \mathrm{D})_z} \sum_{y \in \mathcal{Y}} \sum_{w \in \mathcal{W}} s_w
\end{aligned}$$

$$\sum_{x\in\mathcal{X}} (\pi \triangleright \mathrm{C})_y \, (\pi \triangleright \mathrm{C})_{x|y} \, \mathrm{R}_{y,z} \, g(w,x) \qquad \text{\# Rearrange terms}$$

$$= \frac{1}{(\pi \triangleright \mathrm{D})_z} \sum_{y\in\mathcal{Y}} \sum_{w\in\mathcal{W}} s_w$$

$$\sum_{x\in\mathcal{X}} (\pi \triangleright \mathrm{C})_{x,y} \, \mathrm{R}_{y,z} \, g(w,x) \qquad \text{\# From posterior (6) to joint}$$

$$= \frac{1}{(\pi \triangleright \mathrm{D})_z} \sum_{y\in\mathcal{Y}} \sum_{w\in\mathcal{W}} s_w$$

$$\sum_{x\in\mathcal{X}} \pi_x \, \mathrm{C}_{x,y} \, \mathrm{R}_{y,z} \, g(w,x) \qquad \text{\# From joint (4) to prior \& channel}$$

$$= \frac{1}{(\pi \triangleright \mathrm{D})_z} \sum_{w\in\mathcal{W}} s_w$$

$$\sum_{x\in\mathcal{X}} g(w,x) \, \pi_x \sum_{y\in\mathcal{Y}} \mathrm{C}_{x,y} \, \mathrm{R}_{y,z} \qquad \text{\# Rearrange terms}$$

$$= \frac{1}{(\pi \triangleright \mathrm{D})_z} \sum_{w\in\mathcal{W}} s_w$$

$$\sum_{x\in\mathcal{X}} g(w,x) \, \pi_x \, \mathrm{D}_{x,z} \qquad \text{\# Matrix multiplication CR; \# } \mathrm{D} = \mathrm{C}\,;\mathrm{R} \text{ (Definition 1)}$$

$$= \sum_{w\in\mathcal{W}} s_w \sum_{x\in\mathcal{X}} g(w,x) \, \frac{(\pi \triangleright \mathrm{D})_{x,z}}{(\pi \triangleright \mathrm{D})_z} \qquad \text{\# From (4); \# Rearrange terms}$$

$$= \sum_{w\in\mathcal{W}} s_w \sum_{x\in\mathcal{X}} g(w,x) \, (\pi \triangleright \mathrm{D})_{x|z} \qquad \text{\# From (6)}$$

Since the above holds for any $s \in \mathrm{st}_g((\pi \triangleright \mathrm{D})_{X|z})$, it also holds for the uniform strategy $\mathrm{st}^u_g((\pi \triangleright \mathrm{D})_{X|z})$, which from Defn 9 gives $V^{\mathrm{st}}_g((\pi \triangleright \mathrm{D})_{X|z} \,\|\, (\pi \triangleright \mathrm{D})_{X|z})$ and by Lemma 2 is $V_g((\pi \triangleright \mathrm{D})_{X|z})$. The proof for the second part (ℓ-uncertainty) can be obtained in a similar way, and thus is omitted. □

Theorem 8. *Let $\pi : \mathbb{D}\mathcal{X}$ be the adversary's prior belief, and consider two channels $\mathrm{C} : \mathcal{X} \to \mathbb{D}\mathcal{Y}$ and $\mathrm{D} : \mathcal{X} \to \mathbb{D}\mathcal{Z}$, $\mathrm{C} \sqsubseteq \mathrm{D}$. For any gain function $g : \mathcal{W} \times \mathcal{X} \to \mathbb{R}$, the* expected *posterior g-vulnerability $V_g[\pi \triangleright \mathrm{D}]$ is recovered as*

$$\sum_{z\in\mathcal{Z}} \sum_{y\in\mathcal{Y}} (\pi \triangleright \mathrm{C})_y \, \mathrm{R}_{y,z} \, V^{\mathrm{st}}_g((\pi \triangleright \mathrm{C})_{X|y} \,\|\, (\pi \triangleright \mathrm{D})_{X|z})$$

Dually, for any loss function $\ell : \mathcal{W} \times \mathcal{X} \to \mathbb{R}_{\geq 0} \cup \{\infty\}$, the expected *posterior ℓ-uncertainty $U_\ell[\pi \triangleright \mathrm{D}]$ is recovered as*

$$\sum_{z\in\mathcal{Z}} \sum_{y\in\mathcal{Y}} (\pi \triangleright \mathrm{C})_y \, \mathrm{R}_{y,z} \, U^{\mathrm{st}}_\ell((\pi \triangleright \mathrm{C})_{X|y} \,\|\, (\pi \triangleright \mathrm{D})_{X|z})$$

Proof. This follows almost immediately from Lemma A.2:

$$\sum_{z\in\mathcal{Z}} \sum_{y\in\mathcal{Y}} (\pi \triangleright \mathrm{C})_y \, \mathrm{R}_{y,z} \, V^{\mathrm{st}}_g((\pi \triangleright \mathrm{C})_{X|y} \,\|\, (\pi \triangleright \mathrm{D})_{X|z})$$

$$= \sum_{z\in\mathcal{Z}} (\pi \triangleright \mathrm{D})_z \left(\frac{1}{(\pi \triangleright \mathrm{D})_z} \sum_{y\in\mathcal{Y}} (\pi \triangleright \mathrm{C})_y \, \mathrm{R}_{y,z} \, V^{\mathrm{st}}_g((\pi \triangleright \mathrm{C})_{X|y} \,\|\, (\pi \triangleright \mathrm{D})_{X|z}) \right)$$

$$= \sum_{z\in\mathcal{Z}} (\pi \triangleright \mathrm{D})_z \, V_g((\pi \triangleright \mathrm{D})_{X|z}) \qquad \text{\# Lemma A.2}$$

$$= V_g[\pi \triangleright \mathrm{D}] \qquad \text{\# Definition 5}$$

The proof for the second part (ℓ-uncertainty) can be obtained in a similar way, and thus is omitted. □

Theorem 9. *Let $\pi : \mathbb{D}\mathcal{X}$ be the adversary's prior belief, and consider two channels $\mathrm{C} : \mathcal{X} \to \mathbb{D}\mathcal{Y}$ and $\mathrm{D} : \mathcal{X} \to \mathbb{D}\mathcal{Z}$, $\mathrm{C} \sqsubseteq \mathrm{D}$. For any gain function $g : \mathcal{W} \times \mathcal{X} \to \mathbb{R}$, the* max-case *posterior g-vulnerability $V^{\max}_g[\pi \triangleright \mathrm{D}]$ is recovered as*

$$\max_{z\in\mathcal{Z}} \frac{1}{(\pi \triangleright \mathrm{D})_z} \sum_{y\in\mathcal{Y}} (\pi \triangleright \mathrm{C})_y \, \mathrm{R}_{y,z} \, V^{\mathrm{st}}_g((\pi \triangleright \mathrm{C})_{X|y} \,\|\, (\pi \triangleright \mathrm{D})_{X|z})$$

Dually, for any loss function $\ell : \mathcal{W} \times \mathcal{X} \to \mathbb{R}_{\geq 0} \cup \{\infty\}$, the min-case *posterior ℓ-uncertainty $U^{\min}_\ell[\pi \triangleright \mathrm{D}]$ is recovered as*

$$\min_{z\in\mathcal{Z}} \frac{1}{(\pi \triangleright \mathrm{D})_z} \sum_{y\in\mathcal{Y}} (\pi \triangleright \mathrm{C})_y \, \mathrm{R}_{y,z} \, U^{\mathrm{st}}_\ell((\pi \triangleright \mathrm{C})_{X|y} \,\|\, (\pi \triangleright \mathrm{D})_{X|z})$$

Proof. This follows immediately from Lemma A.2:

$$\max_{z\in\mathcal{Z}} \frac{1}{(\pi \triangleright \mathrm{D})_z} \sum_{y\in\mathcal{Y}} (\pi \triangleright \mathrm{C})_y \, \mathrm{R}_{y,z} \, V^{\mathrm{st}}_g((\pi \triangleright \mathrm{C})_{X|y} \,\|\, (\pi \triangleright \mathrm{D})_{X|z})$$

$$= \max_{z\in\mathcal{Z}} V_g((\pi \triangleright \mathrm{D})_{X|z}) \qquad \text{\# Lemma A.2}$$

$$= V^{\max}_g[\pi \triangleright \mathrm{D}] \qquad \text{\# Definition 5}$$

The proof for the second part (ℓ-uncertainty) can be obtained in a similar way, and thus is omitted. □