Hybrid Statistical Estimation of Mutual Information for Quantifying Information Flow

Yusuke Kawamoto, Fabrizio Biondi, Axel Legay

Hybrid Statistical Estimation of Mutual Information for Quantifying Information Flow

Yusuke Kawamoto[1], Fabrizio Biondi[2], and Axel Legay[2]

[1] AIST, Japan
[2] INRIA, France

Abstract. Analysis of a probabilistic system often requires to learn the joint probability distribution of its random variables. The computation of the exact distribution is usually an exhaustive *precise analysis* on all executions of the system. To avoid the high computational cost of such an exhaustive search, *statistical analysis* has been studied to efficiently obtain approximate estimates by analyzing only a small but representative subset of the system's behavior. In this paper we propose a *hybrid statistical estimation method* that combines precise and statistical analyses to estimate mutual information and its confidence interval. We show how to combine the analyses on different components of the system with different precision to obtain an estimate for the whole system. The new method performs weighted statistical analysis with different sample sizes over different components and dynamically finds their optimal sample sizes. Moreover it can reduce sample sizes by using prior knowledge about systems and a new *abstraction-then-sampling* technique based on qualitative analysis. We show the new method outperforms the state of the art in quantifying information leakage.

1 Introduction

In modeling and analyzing software and hardware systems, the statistical approach is often useful to evaluate quantitative aspects of the behaviors of the systems. In particular, probabilistic systems with complicated internal structures can be approximately and efficiently modeled and analyzed. For instance, statistical model checking has widely been used to verify quantitative properties of many kinds of probabilistic systems [1].

The *statistical analysis* of a probabilistic system is usually considered as a black-box testing approach in which the analyst does not require prior knowledge of the internal structure of the system. The analyst runs the system many times and records the execution traces to construct an approximate model of the system. Even when the formal specification or precise model of the system is not provided to the analyst, statistical analysis can be directly applied to the system if the analyst can execute the black-box implementation. Due to this random sampling of the systems, statistical analysis provides only approximate estimates. However, it can evaluate the accuracy and error of the analysis for instance by providing the confidence intervals of the estimated values.

One of the important challenges in statistical analysis is to estimate entropy-based properties in probabilistic systems. For example, statistical methods [2,3,4,5,6] have been studied for *quantitative information flow analysis* [7,8,9], which estimates an entropy-based property to quantify the leakage of confidential information in a system.

More specifically, the analysis estimates *mutual information* or other properties between two random variables on the secrets and on the observable outputs in the system to measure the amount of information that is inferable about the secret by observing the output. The main technical difficulties in the estimation of entropy-based properties are

1. to efficiently compute large matrices that represent probability distributions, and
2. to provide a statistical method for correcting the bias of the estimate and computing a confidence interval to evaluate the accuracy of the estimation.

To overcome these difficulties we propose a method for statistically estimating mutual information, one of the most popular entropy-based properties. The new method, called *hybrid statistical estimation method*, integrates black-box statistical analysis and white-box *precise analysis*, exploiting the advantages of both. More specifically, this method employs some prior knowledge on the system and performs precise analysis (e.g., static analysis of the source code or specification) on some components of the system. Since precise analysis computes the exact sub-probability distributions of the components, the hybrid method using precise analysis is more accurate than statistical analysis alone.

Moreover, the new method can combine multiple statistical analyses on different components of the system to improve the accuracy and efficiency of the estimation. This is based on our new theoretical results that extend and generalize previous work [10,11,2] on purely statistical estimation. As far as we know this is the first work on a hybrid method for estimating entropy-based properties and their confidence intervals.

To illustrate the method we propose, Fig. 1 presents an example of a joint probability distribution P_{XY} between two random variables X and Y, built up from 3 overlapping components S_1, S_2 and T. To estimate the full joint distribution P_{XY}, the analyst separately computes the joint sub-distribution for the component T by precise analysis, estimates those for S_1 and S_2 by statistical analysis, and then combines these sub-distributions. Since the statistical analysis is based on the random sampling of execution traces, the empirical sub-distributions for S_1 and S_2 are different from the true ones, while the sub-distribution for T is exact. From these approximate and precise sub-distributions, the proposed method can estimate the mutual information for the entire system and evaluate its accuracy by providing a confidence interval. Owing to the combination of different kinds of analyses (with possibly different parameters such as sample sizes), the computation of the bias and confidence interval of the estimate is more complicated than the previous work on statistical analysis.

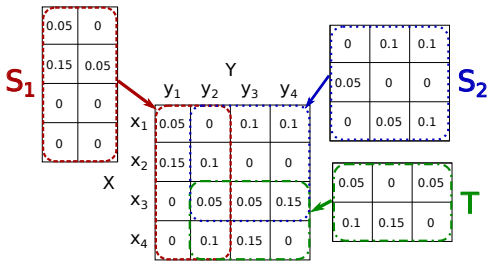


Fig. 1: Joint distribution composed of 3 components.

1.1 Contributions

The contributions of this paper are as follows:

- We propose a new method, called hybrid statistical estimation, that combines statistical and precise analyses on the estimation of mutual information (which can also be applied to Shannon entropy and conditional Shannon entropy). Specifically, we show theoretical results on compositionally computing the bias and confidence interval of the estimate from multiple statistical and precise analyses;
- We present a weighted statistical analysis method with different sample sizes over different components and a method for adaptively optimizing sample sizes for different components by evaluating the quality and cost of the analysis;
- We show how to reduce the sample sizes by using prior knowledge about systems, including an abstraction-then-sampling technique based on qualitative analysis;
- We show that the proposed method can be applied not only to composed systems but also to the source codes of a single system by decomposing it into components and determine the analysis method for each component;
- We evaluate the quality of the estimation in this method, showing that the estimates are more accurate than statistical analysis alone for the same sample size, and that the new method outperforms the state-of-the-art statistical analysis tool LeakWatch [5];
- We demonstrate the effectiveness of the hybrid method in case studies on the quantification of information leakage.

The rest of the paper is structured as follows. Section 2 introduces background in information theory and quantification of information. We compare precise analysis with statistical analysis for the estimation of mutual information. Section 3 describes the main results of this paper: the hybrid method for mutual information estimation, including the method for optimizing sample sizes for different components. Section 4 presents how to reduce sample sizes by using prior knowledge about systems, including the abstraction-then-sampling technique with qualitative analysis. Section 5 overviews how to decompose the source code of a system into components and to determine the analysis method for each component. Section 6 evaluates the proposed method and illustrates its effectiveness against the state of the art. Section 7 discusses related work and Section 8 concludes the paper. All proofs can be found in Appendix A.

2 Information Theory and Quantification of Information

In this section we introduce some background on information theory, which we use to quantify the amount of information in a system. We write X and Y to denote two random variables, and $\mathcal{X}$ and $\mathcal{Y}$ to denote the sets of all possible values of X and Y, respectively. We denote the number of elements of a set $\mathcal{S}$ by $\#\mathcal{S}$.

2.1 Channels

In information theory, a *channel* models the input-output relation of a system as a conditional probability distribution of outputs given inputs. This model has also been used to formalize information leakage in a system that processes confidential data: *inputs* and *outputs* of a channel are respectively regarded as *secrets* and *observables* in the system and the channel represents relationships between the secrets and observables.

A *discrete channel* is a triple $(\mathcal{X}, \mathcal{Y}, C)$ where $\mathcal{X}$ and $\mathcal{Y}$ are two finite sets of discrete input and output values respectively and C is an $\#\mathcal{X} \times \#\mathcal{Y}$ matrix where each element $C[x, y]$ represents the conditional probability of an output y given an input x; i.e., for each $x \in \mathcal{X}$, $\sum_{y \in \mathcal{Y}} C[x, y] = 1$ and $0 \leq C[x, y] \leq 1$ for all $y \in \mathcal{Y}$.

A *prior* is a probability distribution on input values $\mathcal{X}$. Given a prior P_X over $\mathcal{X}$ and a channel C from $\mathcal{X}$ to $\mathcal{Y}$, the *joint probability distribution* P_{XY} of X and Y is defined by: $P_{XY}[x, y] = P_X[x]C[x, y]$ for each $x \in \mathcal{X}$ and $y \in \mathcal{Y}$.

2.2 Mutual Information

The amount of information gained about a random variable X by knowing a random variable Y is defined as the difference between the uncertainty about X before and after observing Y. The *mutual information* $I(X;Y)$ between X and Y is one of the most popular measures to quantify the amount of information on X gained/leaked by Y:

$$I(X;Y) = \sum_{x \in \mathcal{X}, y \in \mathcal{Y}} P_{XY}[x, y] \log_2 \left(\frac{P_{XY}[x, y]}{P_X[x] P_Y[y]} \right)$$

where P_Y is the marginal probability distribution defined as $P_Y[y] = \sum_{x \in \mathcal{X}} P_{XY}[x, y]$.

In the security scenario, information-theoretical measures quantify the amount of secret information leaked against some particular attacker: the mutual information between two random variables X on the secrets and Y on the observables in a system measures the information that is inferable about the secret by knowing the observable. In this scenario mutual information, or Shannon leakage, assumes an attacker that can ask binary questions on the secret's value after observing the system while min-entropy leakage [12] considers an attacker that has only one attempt to guess the secret's value.

Mutual information has been employed in many other applications including Bayesian networks [13], telecommunications [14], pattern recognition [15], machine learning [16], quantum physics [17], and biology [18]. In this work we focus on mutual information for the above security scenario as well as for other purposes such as decision tree training.

2.3 Precise Analysis vs. Statistical Analysis

The calculation of the mutual information $I(X;Y)$ between input X and output Y in a probabilistic system requires the computation of the joint probability distribution P_{XY} of X and Y. The joint distribution can be computed precisely or estimated statistically.

Precise Analysis To obtain the exact joint probability $P_{XY}[x, y]$ for each $x \in \mathcal{X}$ and $y \in \mathcal{Y}$, we sum the probabilities of all execution traces of the system that have input x and output y. This means the computation time depends on the number of traces in the system. If the system has a very large number of traces, it is intractable for analysts to precisely compute the joint distribution and consequently the mutual information.

In [19] the calculation of mutual information is shown to be computationally expensive. This computational difficulty comes from the fact that entropy-based properties are hyperproperties [20] that are defined using all execution traces of the system and

therefore cannot be verified on each single trace. For example, when we investigate the leakage of confidential information in a system, it is insufficient to check the information leakage separately for each component of the system, because the attacker may derive sensitive information by combining the outputs of different components. More generally, the computation of entropy-based properties (such as the amount of leaked information) is not compositional in the sense that an entropy-based property of a system is not the (weighted) sum of those of the components.

For this reason it is inherently difficult to naïvely combine analyses of different components of a system to compute entropy-based properties. In fact, previous studies on the compositional approach in quantitative information flow analysis have faced certain difficulties in obtaining useful bounds on information leakage [21,22,23,24].

Statistical Analysis Due to the complexity of precise analysis, some previous studies have focused on computing approximate values of entropy-based measures. One of the common approaches is the *statistical analysis* based on Monte Carlo methods, in which approximate values are computed from repeated random sampling. Previous work on quantitative information flow has used statistical analysis to mutual information [2,10,11], channel capacity [2,6] and min-entropy leakage [5,25].

In the statistical estimation of mutual information between two random variables X and Y in a probabilistic system, analysts execute the system many times and collect the execution traces each recording a pair of values $(x, y) \in \mathcal{X} \times \mathcal{Y}$. This set of execution traces is used to estimate the empirical joint distribution $\hat{P}_{XY}$ of X and Y and then to compute the mutual information $I(X;Y)$.

Note that the empirical distribution $\hat{P}_{XY}$ is different from the true distribution P_{XY} and thus the estimated mutual information is different from the true value. In fact, it is known that entropy-based measures such as mutual information and min-entropy leakage have some bias and error that depends on the number of collected traces, the matrix size and other factors. However, results on statistics allow us to correct the bias of the estimate and to compute its 95% confidence interval. This way we can guarantee the quality of the estimation, which differentiates our approach from *testing*.

Comparing the Two Analysis Methods The cost of the statistical analysis is proportional to the size $\#\mathcal{X} \times \#\mathcal{Y}$ of the joint distribution matrix (strictly speaking, to the number of non-zero elements in the matrix). Therefore, this method is significantly more efficient than precise analysis if the matrix is relatively small and the number of all traces is very large (for instance because the system's internal variables have a large range). On the other hand, if the matrix is very large, the number of executions needs to be very large to obtain a reliable and small confidence interval. In particular, for a small sample size, statistical analysis does not detect rare events, i.e., traces with a low probability that affect the result.

	Precise	**Statistical**
Type	White box	Black/gray box
Analyzes	Source code	Implementation
Impractical for	Large number of traces	Large matrices
Produces	Exact value	Estimate & confidence

Table 1: Comparison of the two analysis methods.

Main differences between precise and statistical analysis are summarized in Table 1.

3 Hybrid Statistical Estimation of Mutual Information

To overcome the above limitations on the previous approaches we introduce a new method, called *hybrid statistical estimation method*, that integrates both precise and statistical analyses. In this section we present the method for estimating the mutual information between two random variables X (over the inputs $\mathcal{X}$) and Y (over the outputs $\mathcal{Y}$) in a probabilistic system S, and for providing a confidence interval of this estimate. In the method we perform different types of analysis (with different parameters) on different components of a system.

- If a component is deterministic, we perform a precise analysis on it.
- If a component S_i has a joint sub-distribution matrix over *small* subsets of $\mathcal{X}$ and $\mathcal{Y}$ (relatively to the number of all traces), then we perform a statistical analysis on S_i.
- If a component T_j has a *large* matrix (relatively to the number of all traces), we perform a precise analysis on T_j.
- By combining the analysis results on all components we compute the mutual information estimate and its confidence interval. See the rest of Section 3 for details.
- By *qualitative* information flow analysis, the analyst may obtain partial knowledge on components and reduce the sample sizes. See Section 4 for details.

One of the main advantages of the new method is that we guarantee the quality of the outcome by providing its confidence interval even though different kinds of analyses with different parameters are combined together, such as multiple statistical analyses with different sample sizes.

Another advantage is the compositionality in estimating bias and confidence intervals. The random sampling of execution traces is performed independently for each component. Thanks to this we obtain that the bias and confidence interval of mutual information can be computed in a compositional way. This compositionality enables us to find optimal sample sizes for the different components that maximize the accuracy of the estimation (i.e., minimize the confidence interval size) given a fixed total sample size for the entire system. On the other hand, the computation of mutual information itself is not compositional; It requires calculating the *full* joint probability distribution of the system by summing the joint sub-distributions of all components of the system.

Note that these results can be applied to the estimation of Shannon entropy and conditional Shannon entropy as special cases. See Appendix B for the details.

3.1 Computation of Probability Distributions

We consider a probabilistic system S that consists of $(m+k)$ components $S_1, S_2, \ldots, S_m$ and $T_1, T_2, \ldots, T_k$ each executed with probabilities $\theta_1, \theta_2, \ldots, \theta_m$ and $\xi_1, \xi_2, \ldots, \xi_k$; i.e., when S is executed, it yields S_i with the probability θ_i and T_j with the probability ξ_j. We assume S does not have non-deterministic transitions. Let $\mathcal{I} = \{1, 2, \ldots, m\}$ and $\mathcal{J} = \{1, 2, \ldots, k\}$, one of which can be empty. We assume the analyst can run the

component S_i for each $i \in \mathcal{I}$ to record its execution traces, and precisely analyze the components T_j for $j \in \mathcal{J}$, e.g., by static analysis of the source code or specification.

In the estimation of mutual information between two random variables X and Y in the system S, we need to estimate the joint distribution P_{XY} of X and Y. In our approach this is obtained by combining the joint *sub-probability distributions* of X and Y for all the components S_i's and T_j's. More specifically, let R_i and Q_j be the joint sub-distributions of X and Y for the components S_i's and T_j's respectively. Then the joint (full) distribution P_{XY} for the whole system S is defined by:

$$P_{XY}[x,y] \stackrel{\text{def}}{=} \sum_{i \in \mathcal{I}} R_i[x,y] + \sum_{j \in \mathcal{J}} Q_j[x,y]$$

for $x \in \mathcal{X}$ and $y \in \mathcal{Y}$. Note that for each $i \in \mathcal{I}$ and $j \in \mathcal{J}$, the sums of all probabilities in R_i and Q_j equal the probabilities θ_i and ξ_j of executing S_i and T_j respectively.

To estimate the joint distribution P_{XY} the analyst computes

- for each $i \in \mathcal{I}$, the *empirical* sub-distribution $\hat{R}_i$ for the component S_i from a set of traces obtained by executing S_i, and
- for each $j \in \mathcal{J}$, the *exact* sub-distribution Q_j for T_j by a precise analysis on T_j.

The empirical sub-distribution $\hat{R}_i$ is constructed as follows. Let n_i be the number of S_i's executions. For each $x \in \mathcal{X}$ and $y \in \mathcal{Y}$, let K_{ixy} be the number of S_i's traces that have input x and output y. Then $n_i = \sum_{x \in \mathcal{X}, y \in \mathcal{Y}} K_{ixy}$. From these we compute the empirical joint (full) distribution $\hat{D}_i$ of X and Y by $\hat{D}_i[x,y] \stackrel{\text{def}}{=} \frac{K_{ixy}}{n_i}$. Since S_i is executed with probability θ_i, $\hat{R}_i$ is given by $\hat{R}_i[x,y] \stackrel{\text{def}}{=} \theta_i \hat{D}_i[x,y] = \frac{\theta_i K_{ixy}}{n_i}$.

3.2 Estimation of Mutual Information and its Confidence Interval

In this section we present our new method for estimating mutual information and its confidence interval. For each component S_i let D_i be the joint (full) distribution of X and Y obtained by normalizing R_i: $D_i[x,y] = \frac{R_i[x,y]}{\theta_i}$. Let $D_{Xi}[x] = \sum_{y \in \mathcal{Y}} D_i[x,y]$, $D_{Yi}[y] = \sum_{x \in \mathcal{X}} D_i[x,y]$ and $\mathcal{D} = \{(x,y) \in \mathcal{X} \times \mathcal{Y} : P_{XY}[x,y] \neq 0\}$.

Using the estimated $\hat{P}_{XY}$ we can compute the mutual information estimate $\hat{I}(X;Y)$. Note that the mutual information of the whole system is smaller than (or equals) the weighted sum of those of the components, because of its convexity w.r.t. the channel matrix. Therefore it cannot be computed compositionally from those of the components; i.e., it requires to compute the joint distribution matrix $\hat{P}_{XY}$ for the whole system.

Since $\hat{I}(X;Y)$ is obtained from a limited number of traces, it is different from the true value $I(X;Y)$. The following theorem quantifies the bias $E(\hat{I}(X;Y)) - I(X;Y)$.

Theorem 1. *The expectation $E(\hat{I}(X;Y))$ of the mutual information is given by:*

$$I(X;Y) + \sum_{i \in \mathcal{I}} \frac{\theta_i^2}{2n_i} \Bigg(\sum_{(x,y) \in \mathcal{D}} \varphi_{ixy} - \sum_{x \in \mathcal{X}^+} \varphi_{ix} - \sum_{y \in \mathcal{Y}^+} \varphi_{iy} \Bigg) + \mathcal{O}(n_i^{-2})$$

where $\varphi_{ixy} = \frac{D_i[x,y] - D_i[x,y]^2}{P_{XY}[x,y]}$, $\varphi_{ix} = \frac{D_{Xi}[x] - D_{Xi}[x]^2}{P_X[x]}$ and $\varphi_{iy} = \frac{D_{Yi}[y] - D_{Yi}[y]^2}{P_Y[y]}$.

The proof is based on the Taylor expansion w.r.t. multiple dependent variables and can be found in Appendix A. Since the higher-order terms in the formula are negligible when the sample sizes n_i are large enough, we use the following as the *point estimate*:

$$pe = \hat{I}(X;Y) - \sum_{i\in\mathcal{I}} \tfrac{\theta_i^2}{2n_i}\Big(\sum_{(x,y)\in\mathcal{D}} \widehat{\varphi}_{ixy} - \sum_{x\in\mathcal{X}^+} \widehat{\varphi}_{ix} - \sum_{y\in\mathcal{Y}^+} \widehat{\varphi}_{iy}\Big)$$

where $\widehat{\varphi}_{ixy}$, $\widehat{\varphi}_{ix}$ and $\widehat{\varphi}_{iy}$ are empirical values of φ_{ixy}, φ_{ix} and φ_{iy} respectively (that are computed from traces). Then the bias is closer to 0 when the sample sizes n_i are larger.

The quality of the estimate depends on the sample sizes n_i and other factors. The sampling distribution of the estimate $\hat{I}(X;Y)$ tends to follow the normal distribution when n_i's are large enough. The following gives the variance of the distribution.

Theorem 2. *The variance $V(\hat{I}(X;Y))$ of the mutual information is given by*

$$\sum_{i\in\mathcal{I}} \tfrac{\theta_i^2}{n_i}\left(\sum_{(x,y)\in\mathcal{D}} D_i[x,y]\Big(1+\log\tfrac{P_X[x]P_Y[y]}{P_{XY}[x,y]}\Big)^2 - \Big(\sum_{(x,y)\in\mathcal{D}} D_i[x,y]\Big(1+\log\tfrac{P_X[x]P_Y[y]}{P_{XY}[x,y]}\Big)\Big)^2\right) + \mathcal{O}(n_i^{-2})$$

The confidence interval of the estimate of mutual information is useful to know how accurate the estimate is. When the interval is smaller, we learn the estimate is more accurate. The confidence interval is calculated using the variance v obtained by Theorem 2. Given a significance level α, we denote by $z_{\alpha/2}$ the *z-score* for the $100(1-\frac{\alpha}{2})$ percentile point. Then *the* $(1-\alpha)$ *confidence interval* of the estimate is given by:

$$[\,\max(0, pe - z_{\alpha/2}\sqrt{v}),\ \ pe + z_{\alpha/2}\sqrt{v}\,]\ .$$

For example, we use the z-score $z_{0.0025} = 1.96$ to compute the 95% confidence interval. To ignore the higher order terms the sample size $\sum_{i\in\mathcal{I}} n_i$ needs to be at least $4{\cdot}\#\mathcal{X}{\cdot}\#\mathcal{Y}$.

By Theorems 1 and 3, the bias and confidence interval for the whole system can be computed compositionally from those for the components, unlike the mutual information itself. This allows us to adaptively optimize the sample sizes for the components.

3.3 Adaptive Optimization of Sample Sizes

The computational cost of the statistical analysis of each component S_i generally depends on the sample size n_i and the cost of each execution of S_i. When we choose n_i we take into account the trade-off between quality and cost of the analysis: a larger sample size provides a smaller confidence interval, while the cost increases proportionally to n_i.

In this section we present a method for deciding how many times we should run each component S_i to collect a sufficient number of traces to estimate mutual information. More specifically, we show how to compute optimal sample sizes n_i that achieves the smallest confidence interval size within the budget of the total sample size $n = \sum_{i\in\mathcal{I}} n_i$.

To compute the optimal sample sizes, we first run each component to collect a smaller number (for instance dozens) of execution traces. Then we calculate certain intermediate values in computing the variance to determine sample sizes for further executions. Formally, let v_i be the following intermediate value of the variance for S_i:

$$v_i = \theta_i^2\left(\sum_{(x,y)\in\mathcal{D}} \hat{D}_i[x,y]\Big(1+\log\tfrac{\hat{P}_X[x]\hat{P}_Y[y]}{\hat{P}_{XY}[x,y]}\Big)^2 - \Big(\sum_{(x,y)\in\mathcal{D}} \hat{D}_i[x,y]\Big(1+\log\tfrac{\hat{P}_X[x]\hat{P}_Y[y]}{\hat{P}_{XY}[x,y]}\Big)\Big)^2\right)$$

Then we find n_i's that minimize the variance $v = \sum_{i\in\mathcal{I}} \frac{v_i}{n_i}$ of the mutual information.

Theorem 3. *Given the total sample size n and the above intermediate variance v_i of the component S_i for each $i \in \mathcal{I}$, the variance of the mutual information estimate is minimized if, for all $i \in \mathcal{I}$, the sample size n_i for S_i satisfies $n_i = \frac{\sqrt{v_i}n}{\sum_{j=1}^{m}\sqrt{v_j}}$.*

By this result the estimation of a confidence interval size is useful to optimally assign sample sizes to components even when the analyst is not interested in the interval itself. We show experimentally the effectiveness of this optimization in Appendix F.

4 Estimation Using Prior Knowledge about Systems

In this section we show how to use prior knowledge about systems to improve the estimation, i.e., to make the size of the confidence intervals smaller and reduce the required sample sizes.

4.1 Approximate Estimation Using Knowledge of Prior Distributions

Our hybrid statistical estimation method integrates both precise and statistical analysis, and it can be seen as a generalization and extension of previous work [2,10,11].

For example, Chatzikokolakis et.al. [2] present a method for estimating mutual information between two random variables X (over secret values $\mathcal{X}$) and Y (over observable values $\mathcal{Y}$) when the analyst knows the (prior) distribution P_X of X. In the estimation they collect execution traces by running a system for each secret value $x \in \mathcal{X}$. Thanks to the precise knowledge of P_X, they have more accurate estimates than the other previous work [10,11] that also estimates P_X from execution traces.

Estimation using the precise knowledge of P_X is an instance of our result if a system is partitioned into the component S_x for each secret $x \in \mathcal{X} = \mathcal{I}$. If we assume all joint probabilities are non-zero, the approximate result in [2] follows from Theorem 1.

Corollary 1. *The expectation $E(\hat{I}(X;Y))$ of the mutual information is given by*

$$I(X;Y) + \frac{(\#\mathcal{X}-1)(\#\mathcal{Y}-1)}{2n} + \mathcal{O}(n^{-2}).$$

In this result from [2] the bias $\frac{(\#\mathcal{X}-1)(\#\mathcal{Y}-1)}{2n}$ depends only on the size of the joint distribution matrix. However, the bias can be strongly influenced by zeroes or very small probabilities in the distribution, therefore their approximate results can be correct only when all joint probabilities are non-zero and large enough, which is a strong restriction in practice. We show in Appendix E that the tool LeakWatch [5] implicitly assumes that all probabilities are large enough, and consequently miscalculates bias and gives an estimate far from the true value in the presence of very small probabilities.

4.2 Our Estimation Using Knowledge of Prior Distributions

To overcome these issues we present more general results in the case the analyst knows the prior distribution P_X. We assume that a system S is partitioned into the disjoint component S_{ix} for each index $i \in \mathcal{I}$ and secret $x \in \mathcal{X}$, and that each S_{ix} is executed with probability θ_{ix} in the system S. Let $\Theta = \{\theta_{ix} : i \in \mathcal{I},\ x \in \mathcal{X}\}$.

In the estimation of mutual information we run each component S_{ix} separately many times to collect execution traces. Unlike the previous work we may change the number of executions $n_i P_X[x]$ to $n_i \lambda_i[x]$ where $\lambda_i[x]$ is an *importance prior* that decides how the sample size n_i is allocated for each component S_{ix}. Let $\Lambda = \{\lambda_i : i \in \mathcal{I}\}$.

Given the number K_{ixy} of S_{ix}'s traces with output y, we define the conditional distribution D_i of output given input: $D_i[y|x] \stackrel{\text{def}}{=} \frac{K_{ixy}}{n_i \lambda_i[x]}$. Let $M_{ixy} = \frac{\theta_{ix}^2}{\lambda_i[x]} D_i[y|x]\,(1 - D_i[y|x])$. Then the following is the expectation and variance of the mutual information $\hat{I}_{\Theta,\Lambda}(X;Y)$ calculated using $\hat{D}_i$, Θ, Λ.

Proposition 1. *The expectation $E(\hat{I}_{\Theta,\Lambda}(X;Y))$ of the mutual information is given by*

$$I(X;Y) + \sum_{i \in \mathcal{I}} \frac{1}{2n_i} \sum_{y \in \mathcal{Y}^+} \left(\sum_{x \in \mathcal{D}_y} \frac{M_{ixy}}{P_{XY}[x,y]} - \frac{\sum_{x \in \mathcal{D}_y} M_{ixy}}{P_Y[y]} \right) + \mathcal{O}(n_i^{-2})$$

Proposition 2. *The variance $V(\hat{I}_{\Theta,\Lambda}(X;Y))$ of the mutual information is given by*

$$\sum_{i \in \mathcal{I}} \sum_{x \in \mathcal{X}^+} \frac{\theta_{ix}^2}{n_i \lambda_i[x]} \left(\sum_{y \in \mathcal{D}_x} D_i[y|x] \left(\log \frac{P_Y[y]}{P_{XY}[x,y]} \right)^2 - \left(\sum_{y \in \mathcal{D}_x} D_i[y|x] \left(\log \frac{P_Y[y]}{P_{XY}[x,y]} \right) \right)^2 \right) + \mathcal{O}(n_i^{-2})$$

By applying Theorem 3, the sample sizes n_i and the importance priors λ_i can be adaptively optimized. We describe this in the Appendix A.3 due to space constraints.

4.3 Abstraction-Then-Sampling Using Partial Knowledge of Components

In this section we extend our estimation method to consider the case in which the analyst has partial knowledge of components (e.g. by static analysis of the source code or specification) before sampling. Such prior knowledge may help us abstract components into simpler ones and thus reduce the sample size for the statistical analysis.

For instance, let us consider an analyst who knows two pairs (x, y) and (x', y') of inputs and outputs have the same probability in a component S_i: $D_i[x, y] = D_i[x', y']$. Then, when we construct the empirical distribution $\hat{D}_i$ from a set of traces, we can count the number $K_{i\{(x,y),(x',y')\}}$ of traces having either (x, y) or (x', y'), and divide it by two: $K_{ixy} = K_{ix'y'} = \frac{K_{i\{(x,y),(x',y')\}}}{2}$. Then the sample size required for a certain accuracy is smaller than when we do not use the prior knowledge on the equality $K_{ixy} = K_{ix'y'}$.

In the following we generalize this idea to deal with more knowledge of components. Let us consider a (probabilistic) system in which some components leak no information on inputs and the analyst can learn this by *qualitative* information analysis (for verifying non-interference). Then such a component S_i has a sub-channel matrix where all non-zero rows have an identical conditional distribution of outputs given inputs [26]. Consequently, when we estimate the $\#\mathcal{X}_i \times \#\mathcal{Y}_i$ matrix of S_i it suffices to estimate one of the rows, hence the number of executions is proportional to $\#\mathcal{Y}_i$ instead of $\#\mathcal{X}_i \times \#\mathcal{Y}_i$. Note that even when some components leak no information, computing the mutual information for the whole system requires constructing the matrix of the system, hence the matrices of all components.

The following results show that the bias and confidence interval are narrower than when not using the prior knowledge of components. Let $\mathcal{I}^\star$ be the set of indices of components that have channel matrices whose non-zero rows consist of the same distribution. For each $i \in \mathcal{I}^\star$, we define $\pi_i[x]$ as the probability of having an input x in the component S_i. Then the expectation and variance of the mutual information are as follows.

Theorem 4. *The expectation $E(\hat{I}_{\mathcal{I}^\star}(X;Y))$ of the mutual information is given by*

$$I(X;Y)+\sum_{i\in\mathcal{I}\setminus\mathcal{I}^\star}\frac{\theta_i^2}{2n_i}\Big(\sum_{(x,y)\in\mathcal{D}}\varphi_{ixy}-\sum_{x\in\mathcal{X}^+}\varphi_{ix}-\sum_{y\in\mathcal{Y}^+}\varphi_{iy}\Big)+\sum_{i\in\mathcal{I}^\star}\frac{\theta_i^2}{2n_i}\Big(\sum_{(x,y)\in\mathcal{D}}\psi_{ixy}-\sum_{y\in\mathcal{Y}^+}\varphi_{iy}\Big)+\mathcal{O}(n_i^{-2})$$

where $\psi_{ixy} \stackrel{\text{def}}{=} \frac{D_i[x,y]\pi_i[x]-D_i[x,y]^2}{P_{XY}[x,y]}$.

Theorem 5. *The variance $V(\hat{I}_{\mathcal{I}^\star}(X;Y))$ of the mutual information is given by*

$$\sum_{i\in\mathcal{I}\setminus\mathcal{I}^\star}\frac{\theta_i^2}{n_i}\Bigg(\sum_{(x,y)\in\mathcal{D}}D_i[x,y]\Big(1+\log\tfrac{P_X[x]P_Y[y]}{P_{XY}[x,y]}\Big)^2-\Big(\sum_{(x,y)\in\mathcal{D}}D_i[x,y]\Big(1+\log\tfrac{P_X[x]P_Y[y]}{P_{XY}[x,y]}\Big)\Big)^2\Bigg)$$
$$+\sum_{i\in\mathcal{I}^\star}\frac{\theta_i^2}{n_i}\Bigg(\sum_{y\in\mathcal{Y}^+}D_{Yi}[y]\Big(\log P_Y[y]-\sum_{x\in\mathcal{X}}\pi_i[x]\log P_{XY}[x,y]\Big)^2$$
$$-\Big(\sum_{y\in\mathcal{Y}^+}D_{Yi}[y]\Big(\log P_Y[y]-\sum_{x\in\mathcal{X}}\pi_i[x]\log P_{XY}[x,y]\Big)\Big)^2\Bigg)+\mathcal{O}(n_i^{-2})\,.$$

5 Estimation via Program Decomposition

The hybrid statistical estimation presented in the previous sections is designed to analyze a system composed of subsystems (for instance, a distributed system over different software or hardware, potentially geographically separated). However, it can also be applied to the the source code of a system by decomposing it into disjoint components. In this section we show how to decompose a code into components and determine for each component which analysis method to use and the method's parameters.

The principles to decompose a system's source code in components are as follows:

- The code may be decomposed only at conditional branching. Moreover, each component must be a terminal in the control flow graph, hence no component is executed afterwards. This is because the estimation method requires that the channel matrix for the system is the weighted sum of those for its components, and that the weight of a component is the probability of executing it.
- The analysis method and its parameters for each component S_i are decided by estimating the computational cost of analyzing S_i. Let $\mathcal{Z}_i$ be the set of all *internal randomness* (i.e., the variables whose values are assigned according to probability distributions) in S_i. Then the cost of the statistical analysis is proportional to S_i's matrix size $\#\mathcal{X}_i \times \#\mathcal{Y}_i$, while the cost of the precise analysis is proportional to the number of all traces in S_i's control flow graph (in the worst case proportional to $\#\mathcal{X}_i \times \#\mathcal{Z}_i$). Hence the cost estimation is reduced to counting $\#\mathcal{Y}_i$ and $\#\mathcal{Z}_i$.

The procedure for decomposition is shown in Fig. 2 and is illustrated in Appendix D using the example of Fig. 7. Since this is heuristic, it is not guaranteed to produce an optimal decomposition. While the procedure is automated, for usability the choice of analysis can be controlled by user's annotations on the code.

1. Build the control flow graph of the system.
2. Mark all possible components based on each conditional branching. Each possible component must be a terminal as explained in Section 5.
3. For each possible component S_i, check whether it is deterministic or not (by syntactically checking an occurrence of a probabilistic assignment or a probabilistic function call). If it is, mark the component for precise analysis.
4. For each possible component S_i, check whether S_i's output variables are independent of its input variables inside S_i (by *qualitative* information flow). If so, mark that the abstraction-then-sampling technique in Section 4.3 is to be used on the component.
5. For each S_i, estimate an approximate range size of its internal and observable variables.
6. Looking from the leaves to the root of the graph, decide the decomposition into components. Estimate the cost of statistical and precise analyses and mark the component for analysis by the cheapest of the two.
7. Join together adjacent components if they are marked for precise analysis, or if they are marked for statistical analysis and have the same input and output ranges.
8. For each component, perform precise analysis or statistical analysis as marked.

Fig. 2: Procedure for decomposing a system given its source code.

6 Evaluation

We evaluate experimentally the effectiveness of our hybrid method compared to the state of the art. We first discuss the cost and quality of the estimation, then test the hybrid method against fully precise/fully statistical analyses on Shannon leakage benchmarks. Another case study (on decision tree training) is shown in Appendix F.

6.1 On the Tradeoff between the Cost and Quality of Estimation

In the hybrid statistical estimation, the estimate takes different values probabilistically, because it is computed from a set of traces that are generated by executing a probabilistic system. Fig. 3 shows the sampling distribution of the mutual information estimate of the joint distribution in Fig. 1 in Section 1. The graph shows the frequency (on the y axis) of the mutual information estimates (on the x axis) when performing the estimation 1000 times. In each estimation we perform precise analysis on the component T and statistical analysis on S_1 and S_2 (with a sample size of 5000). Details of the experiments are provided in Appendix C. As shown in Fig. 3 the

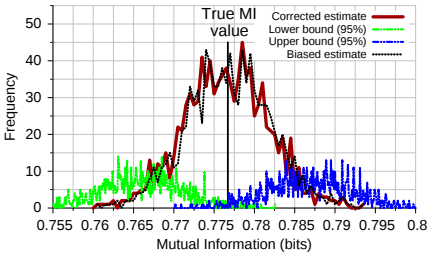
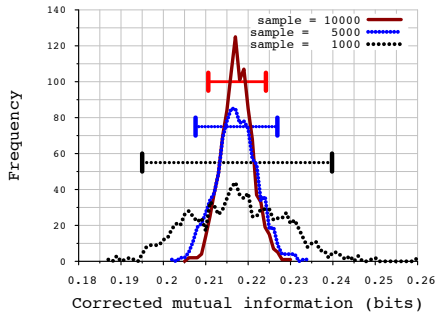
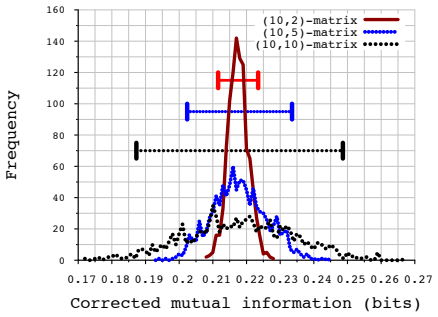


Fig. 3: Distribution of mutual information estimate and its confidence interval.



(a) Estimates and sample sizes.



(b) Estimates and the ratio of precise analysis.

Fig. 4: Smaller intervals when increasing the sample size or the ratio of precise analysis.

estimate after the correction of bias by Theorem 1 is closer to the true value. The estimate is roughly between the lower and upper bounds of the 95% confidence interval calculated using Theorem 2.

The interval size depends on the sample size in statistical analysis as shown in Fig. 4a. When the sample size is k times larger, the confidence interval is $\sqrt{k}$ times narrower. The interval size also depends on the amount of precise analysis as shown in Fig. 4b. If we perform precise analysis on larger components, then the sampling distribution becomes more centered (with shorter tails) and the confidence interval becomes narrower.

The hybrid approach produces better estimates than the state of the art in statistical analysis. Due to the combination with precise analysis, the confidence interval estimated by our approach is smaller than LeakWatch [5] for the same sample size.

6.2 Shannon Leakage Benchmarks

We compare the performance of our hybrid method with fully precise/statistical analysis on Shannon leakage benchmarks. Our implementations of precise and statistical analyses are variants of the state-of-the art tools QUAIL [27,28] and LeakWatch [5,29] respectively. They are fully automated except for human-provided annotations to determine the analysis method for each component. All experiments are performed on an Intel i7-4960HQ 2.6GHz quad-core machine with 8GB of RAM running Ubuntu 16.04 .

```
1 secret array bit[N] s;
2 observable array bit[K] r;
3 for i=0..K-1 do r[i]=s[i] ;
4 for i=K..N-1 do
5 |   j = uniform(0..i);
6 |   if j<K then r[j]=s[i];
7 end
```

Fig. 5: Reservoir sampling.

Reservoir Sampling The reservoir sampling problem [30] consists of selecting K elements randomly from a pool of $N > K$ elements. We quantify the information flow of the commonly-used *Algorithm R* [30], presented in Fig 5, for various values of N and $K = N/2$. In the algorithm, the first K elements are chosen as the sample, then each other element has a probability to replace one element in the sample.

```
1  secret int h = [0, N];
2  observable array bit[N] decl;
3  int lie = uniform(1..N);
4  randomly generated array bit[N] coin;
5  for c in coin do  c = uniform(0..1) ;
6  for i=0..N-1 do
7  |  decl[i]=coin[i] xor coin[(i+1)%N];
8  |  if h==i+1 then decl[i]=!decl[i];
9  |  if i==lie then decl[i]=!decl[i];
10 end
```

Fig. 6: Lying cryptographers.

Multiple Lying Cryptographers Protocol We test our hybrid method to compute the Shannon leakage of a distributed version of the lying cryptographers protocol. The lying cryptographers protocol is a variant of the dining cryptographer multiparty computation protocol [31] in which a randomly-chosen cryptographer declares the opposite of what they would normally declare, i.e. they lie if they are not the payer, and do not lie if they are the payer. We consider three simultaneous lying cryptographers implementation in which 8 cryptographers run the protocol on three separate overlapping tables A, B and C with 4 cryptographers each. Table A hosts cryptographers 1 to 4, Table B hosts cryptographers 3 to 6, and Table C hosts cryptographers 5 to 8. The identity of the payer is the same in all tables. We discuss this example in more details in Appendix E.

```
1  secret int sec = [0,N-1];
2  observable int obs;
3  int S = uniform(0,N-W-1);
4  int ws = uniform(1,W);
5  int O = uniform(0,N-W-1);
6  int wo = uniform(1,W);
7  if S ≤ sec ≤ S+ws then
8  |  obs = uniform(O,O+wo);
9  else
10 |  obs = uniform(0,N-1);
11 end
```

Fig. 7: Shifting Window.

Shifting Window In the shifting window example the secret has N possible values, and a contiguous sequence of this values (the "window") of random size from 1 to W is chosen. We assume for simplicity that $N = 2W$. If the secret is inside the window then another random window is chosen in the same way and a random value from the new window is printed. Otherwise, a random value from 0 to $N - 1$ is printed.

Results In Table 2 we show the results of the benchmarks using fully precise, fully statistical and hybrid analyses, for a sample size of 100000 executions. Timeout is set at 10 minutes. On the reservoir benchmark the precise analysis is faster for small instances but does not scale, timing out on larger values of N. The hybrid method is consistently faster than the fully statistical analy-

		Reservoir				Lying Crypt	Window		
		N=6	**N=8**	**N=10**	**N=12**		**N=20**	**N=22**	**N=24**
Precise	**Time(s)**	0.7	11.4	timeout	timeout	506.4	10.0	16.0	28.3
	Error	0	0	-	-	0	0	0	0
Statistical	**Time(s)**	21.6	35.2	60.7	91.5	254.3	7.5	7.7	7.1
	Error	10^{-3}	10^{-3}	-	-	10^{-3}	10^{-3}	10^{-3}	10^{-4}
Hybrid	**Time(s)**	13.4	22.5	34.6	58.4	240.1	6.6	7.1	7.1
	Error	10^{-4}	10^{-3}	-	-	10^{-3}	10^{-7}	10^{-4}	10^{-4}

Table 2: Shannon leakage benchmark results.

sis and often has a smaller error. On the other benchmarks the hybrid method usually outperforms the others and produces better approximations than the statistical analysis.

The results in Table 2 show the superiority of our hybrid approach compared to the state of the art. The hybrid analysis scales better than the precise analysis, since it does not need to analyze every trace of the system. Compared to fully statistical analysis, our hybrid analysis exploits precise analysis on components of the system where statistical estimation would be more expensive than precise analysis. This allows the hybrid analysis to focus the statistical estimation on components of the system where it converges faster, thus obtaining a smaller confidence interval in a shorter time.

7 Related Work

The information-theoretical approach to program security dates back to the work of Denning [32] and Gray [33]. Clark et al. [7,34] presented techniques to automatically compute mutual information of an imperative language with loops. For a deterministic program, leakage can be computed from the equivalence relations on the secret induced by the possible outputs, and such relations can be automatically quantified [35]. Under- and over-approximation of leakage based on the observation of some traces have been studied for deterministic programs [36,37]. The combination of static and statistical approaches to quantitative information flow is proposed in [38] while our paper is general enough to deal with probabilistic systems under various prior information conditions.

The statistical approach to quantifying information leakage has been studied since the seminal work by Chatzikokolakis et al. [2]. Chothia et al. have developed this approach in tools leakiEst [3,39] and LeakWatch [5,29]. The hybrid statistical method in this paper can be considered as their extension with the inclusion of component weighting and adaptive priors inspired by the importance sampling in statistical model checking [40,41]. To the best of our knowledge, no prior work has applied weighted statistical analysis to the estimation of mutual information or any other leakage measures.

Fremont and Seshia [42] have presented a polynomial time algorithm to approximate the weight of traces of deterministic programs with possible application to quantitative information leakage. Progress in statistical program analysis includes a scalable algorithm for uniform generation of sample from a distribution defined as constraints [43,44], with applications to constrained-random program verification.

The algorithms for precise computation of information leakage used in this paper are based on trace analysis [45], implemented in the QUAIL tool [28,27]. Phan et al. [46,47] developed tools to compute channel capacity of deterministic programs written in the C or Java languages. McCamant et al. [48] developed tools implementing dynamic quantitative taint analysis techniques for security. The recent tool Moped-QLeak [49] is able to efficiently compute information leakage of programs as long as it can produce a complete symbolic representation of the program.

8 Conclusions and Future Work

We have proposed a method for estimating mutual information by combining precise and statistical analyses and for compositionally computing the bias and confidence interval

of the estimate. The results are also used to adaptively find the optimal sample sizes for different components in the statistical analysis. Moreover, we have shown how to reduce sample sizes by using prior knowledge about systems, including the abstraction-then-sampling technique with qualitative analysis. To apply our new method to the source codes of systems we have shown how to decompose the codes into components and determine the analysis method for each component. We have shown both theoretical and experimental results to demonstrate that the proposed approach outperforms the state of the art. To obtain better results we are developing theory and tools that integrate symbolic abstraction techniques in program analysis into our estimation method.

References

1. Legay, A., Delahaye, B., Bensalem, S.: Statistical model checking: An overview. In: Runtime Verification - First International Conference, RV 2010, St. Julians, Malta, November 1-4, 2010. Proceedings. (2010) 122–135
2. Chatzikokolakis, K., Chothia, T., Guha, A.: Statistical measurement of information leakage. In Esparza, J., Majumdar, R., eds.: TACAS 2010. Proceedings. Volume 6015 of Lecture Notes in Computer Science., Springer (2010) 390–404
3. Chothia, T., Kawamoto, Y., Novakovic, C.: A tool for estimating information leakage. In Sharygina, N., Veith, H., eds.: CAV 2013. Proceedings. Volume 8044 of Lecture Notes in Computer Science., Springer (2013) 690–695
4. Chothia, T., Kawamoto, Y., Novakovic, C., Parker, D.: Probabilistic point-to-point information leakage. In: CSF 2013. Proceedings, IEEE (2013) 193–205
5. Chothia, T., Kawamoto, Y., Novakovic, C.: Leakwatch: Estimating information leakage from java programs. In Kutylowski, M., Vaidya, J., eds.: ESORICS 2014. Proceedings, Part II. Volume 8713 of Lecture Notes in Computer Science., Springer (2014) 219–236
6. Boreale, M., Paolini, M.: On formally bounding information leakage by statistical estimation. In Chow, S.S.M., Camenisch, J., Hui, L.C.K., Yiu, S., eds.: ISC 2014. Proceedings. Volume 8783 of Lecture Notes in Computer Science., Springer (2014) 216–236
7. Clark, D., Hunt, S., Malacaria, P.: Quantitative analysis of the leakage of confidential data. Electr. Notes Theor. Comput. Sci. **59**(3) (2001) 238–251
8. Köpf, B., Basin, D.A.: An information-theoretic model for adaptive side-channel attacks. In: Proc. of CCS, ACM (2007) 286–296
9. Chatzikokolakis, K., Palamidessi, C., Panangaden, P.: Anonymity protocols as noisy channels. Inf. and Comp. **206**(2–4) (2008) 378–401
10. Moddemeijer, R.: On estimation of entropy and mutual information of continuous distributions. Signal Processing **16** (1989) 233–248
11. Brillinger, D.R.: Some data analysis using mutual information. Brazilian Journal of Probability and Statistics **18**(6) (2004) 163–183
12. Smith, G.: On the foundations of quantitative information flow. In de Alfaro, L., ed.: FOSSACS 2009. Proceedings. Volume 5504 of Lecture Notes in Computer Science., Springer (2009) 288–302
13. Jensen, F.V.: Introduction to Bayesian Networks. 1st edn. Springer-Verlag New York, Inc., Secaucus, NJ, USA (1996)
14. Gallager, R.G.: Information Theory and Reliable Communication. John Wiley & Sons, Inc., New York, NY, USA (1968)
15. Escolano, F., Suau, P., Bonev, B.: Information Theory in Computer Vision and Pattern Recognition. Springer, Londres (2009)

16. MacKay, D.J.C.: Information Theory, Inference & Learning Algorithms. Cambridge University Press, New York, NY, USA (2002)
17. Wilde, M.M.: Quantum Information Theory. 1st edn. Cambridge University Press, New York, NY, USA (2013)
18. Adami, C.: Information theory in molecular biology. Physics of Life Reviews **1**(1) (April 2004) 3–22
19. Yasuoka, H., Terauchi, T.: Quantitative information flow as safety and liveness hyperproperties. Theor. Comput. Sci. **538** (2014) 167–182
20. Clarkson, M.R., Schneider, F.B.: Hyperproperties. Journal of Computer Security **18**(6) (2010) 1157–1210
21. Barthe, G., Köpf, B.: Information-theoretic bounds for differentially private mechanisms. In: Proc. of CSF, IEEE (2011) 191–204
22. Espinoza, B., Smith, G.: Min-entropy as a resource. Inf. Comput. (2013)
23. Kawamoto, Y., Chatzikokolakis, K., Palamidessi, C.: Compositionality results for quantitative information flow. In: QEST 2014. Proceedings. (2014) 368–383
24. Kawamoto, Y., Given-Wilson, T.: Quantitative information flow for scheduler-dependent systems. In: QAPL 2015. Proceedings. Volume 194. (2015) 48–62
25. Chothia, T., Kawamoto, Y.: Statistical estimation of min-entropy leakage (April 2014) Manuscript.
26. Cover, T.M., Thomas, J.A.: Elements of information theory (2. ed.). A Wiley-Interscience publication. Wiley (2006)
27. Biondi, F., Legay, A., Traonouez, L., Wasowski, A.: QUAIL: A quantitative security analyzer for imperative code. In Sharygina, N., Veith, H., eds.: CAV 2013. Proceedings. Volume 8044 of Lecture Notes in Computer Science., Springer (2013) 702–707
28. Biondi, F., Legay, A., Traonouez, L.M., Wasowski, A.: QUAIL https://project.inria.fr/quail/.
29. Chothia, T., Kawamoto, Y., Novakovic, C.: LeakWatch http://www.cs.bham.ac.uk/research/projects/infotools/leakwatch/.
30. Vitter, J.S.: Random sampling with a reservoir. ACM Trans. Math. Softw. **11**(1) (March 1985) 37–57
31. Chaum, D.: The dining cryptographers problem: Unconditional sender and recipient untraceability. Journal of Cryptology **1** (1988) 65–75
32. Denning, D.E.: A lattice model of secure information flow. Commun. ACM **19**(5) (1976) 236–243
33. Gray, J.W.: Toward a mathematical foundation for information flow security. In: IEEE Symposium on Security and Privacy. (1991) 21–35
34. Clark, D., Hunt, S., Malacaria, P.: A static analysis for quantifying information flow in a simple imperative language. Journal of Computer Security **15**(3) (2007) 321–371
35. Backes, M., Köpf, B., Rybalchenko, A.: Automatic discovery and quantification of information leaks. In: 30th IEEE Symposium on Security and Privacy (S&P 2009), 17-20 May 2009, Oakland, California, USA, IEEE Computer Society (2009) 141–153
36. McCamant, S., Ernst, M.D.: Quantitative information flow as network flow capacity. In Gupta, R., Amarasinghe, S.P., eds.: Proceedings of the ACM SIGPLAN 2008 Conference on Programming Language Design and Implementation, Tucson, AZ, USA, June 7-13, 2008, ACM (2008) 193–205
37. Newsome, J., McCamant, S., Song, D.: Measuring channel capacity to distinguish undue influence. In Chong, S., Naumann, D.A., eds.: Proceedings of the 2009 Workshop on Programming Languages and Analysis for Security, PLAS 2009, Dublin, Ireland, 15-21 June, 2009, ACM (2009) 73–85
38. Köpf, B., Rybalchenko, A.: Approximation and randomization for quantitative information-flow analysis. In: CSF 2010. Proceedings, IEEE Computer Society (2010) 3–14

39. Chothia, T., Kawamoto, Y., Novakovic, C.: leakiEst http://www.cs.bham.ac.uk/research/projects/infotools/leakiest/.
40. Barbot, B., Haddad, S., Picaronny, C.: Coupling and importance sampling for statistical model checking. In Flanagan, C., König, B., eds.: TACAS 2012. Proceedings. Volume 7214 of Lecture Notes in Computer Science., Springer (2012) 331–346
41. Clarke, E.M., Zuliani, P.: Statistical model checking for cyber-physical systems. In Bultan, T., Hsiung, P., eds.: ATVA 2011. Proceedings. Volume 6996 of Lecture Notes in Computer Science., Springer (2011) 1–12
42. Fremont, D.J., Seshia, S.A.: Speeding up SMT-based quantitative program analysis. In Rümmer, P., Wintersteiger, C.M., eds.: SMT 2014. Proceedings. Volume 1163 of CEUR Workshop Proceedings., CEUR-WS.org (2014) 3–13
43. Chakraborty, S., Fremont, D.J., Meel, K.S., Seshia, S.A., Vardi, M.Y.: On parallel scalable uniform SAT witness generation. In Baier, C., Tinelli, C., eds.: TACAS 2015. Proceedings. Volume 9035 of Lecture Notes in Computer Science., Springer (2015) 304–319
44. Chakraborty, S., Meel, K.S., Vardi, M.Y.: A scalable approximate model counter. In Schulte, C., ed.: CP 2013. Proceedings. Volume 8124 of Lecture Notes in Computer Science., Springer (2013) 200–216
45. Biondi, F., Legay, A., Malacaria, P., Wasowski, A.: Quantifying information leakage of randomized protocols. Theor. Comput. Sci. **597** (2015) 62–87
46. Phan, Q., Malacaria, P.: Abstract model counting: a novel approach for quantification of information leaks. In Moriai, S., Jaeger, T., Sakurai, K., eds.: AsiaCCS 2014. Proceedings, ACM (2014) 283–292
47. Phan, Q., Malacaria, P., Pasareanu, C.S., d'Amorim, M.: Quantifying information leaks using reliability analysis. In Rungta, N., Tkachuk, O., eds.: SPIN 2014. Proceedings, ACM (2014) 105–108
48. Kang, M.G., McCamant, S., Poosankam, P., Song, D.: DTA++: dynamic taint analysis with targeted control-flow propagation. In: NDSS 2011. Proceedings. (2011)
49. Chadha, R., Mathur, U., Schwoon, S.: Computing information flow using symbolic model-checking. In: FSTTCS 2014. Proceedings. (2014) 505–516
50. Biondi, F., Legay, A., Quilbeuf, J.: Comparative analysis of leakage tools on scalable case studies. In Fischer, B., Geldenhuys, J., eds.: SPIN 2015, Proceedings. Volume 9232 of Lecture Notes in Computer Science., Springer (2015) 263–281
51. Quinlan, J.R.: Induction of decision trees. Machine Learning **1**(1) 81–106

A Omitted Proofs

We present proofs omitted in the paper. We refer to literature [26] for the definitions of the variance $V(X)$ and the covariance $Cov(X, Y)$.

Hereafter we denote by Q the joint sub-distribution obtained by summing Q_j's:

$$Q[x, y] \stackrel{\text{def}}{=} \sum_{j \in \mathcal{J}} Q_j[x, y] \quad .$$

We write q_{xy} to denote $Q[x, y]$ for abbreviation. Then q_{xy} is the probability that the execution of S yields one of T_j's and has input x and output y.

A.1 Proofs for the Main Results (without assuming the knowledge of the prior)

Proof (Proof of Theorem 1). Recall that the empirical joint probability is given by

$$\hat{P}_{XY}[x, y] = q_{xy} + \sum_{i \in \mathcal{I}} \frac{\theta_i K_{ixy}}{n_i}.$$

Then the empirical joint entropy is defined by

$$\begin{aligned}\hat{H}(X, Y) &= -\sum_{x \in \mathcal{X}, y \in \mathcal{Y}} \hat{P}_{XY}[x, y] \log \hat{P}_{XY}[x, y] \\ &= -\sum_{x \in \mathcal{X}, y \in \mathcal{Y}} \left(q_{xy} + \sum_{i \in \mathcal{I}} \frac{\theta_i K_{ixy}}{n_i} \right) \log \left(q_{xy} + \sum_{i \in \mathcal{I}} \frac{\theta_i K_{ixy}}{n_i} \right).\end{aligned}$$

We define the set of pairs consisting of secrets and observables that appear with non-zero probabilities in the execution of the whole system S:

$$\mathcal{D} = \{(x, y) \in \mathcal{X} \times \mathcal{Y} : P_{XY}[x, y] \neq 0\}.$$

Let $\overline{K_{ixy}} = E(K_{ixy})$. Then $\overline{K_{ixy}} = n_i D_i[x, y] = \frac{n_i R_i[x,y]}{\theta_i}$. By the Taylor expansion w.r.t. $\overline{K_{ixy}}$ for all $i \in \mathcal{I}$,

$$\begin{aligned}\hat{H}(X, Y) = -\sum_{(x,y) \in \mathcal{D}} (&A_{xy} + \sum_{i \in \mathcal{I}} B_{ixy}(K_{ixy} - \overline{K_{ixy}}) \\ &+ \sum_{i,j \in \mathcal{I},\, i \neq j} C_{ijxy}(K_{ixy} - \overline{K_{ixy}})(K_{jxy} - \overline{K_{jxy}}) \\ &+ \sum_{i \in \mathcal{I}} C_{iixy}(K_{ixy} - \overline{K_{ixy}})^2 + \mathcal{O}(n_i^{-2}))\end{aligned}$$

where

- $A_{xy} = \left(q_{xy} + \sum_{l \in \mathcal{I}} \frac{\theta_l \overline{K_{lxy}}}{n_l} \right) \log \left(q_{xy} + \sum_{l \in \mathcal{I}} \frac{\theta_l \overline{K_{lxy}}}{n_l} \right),$
- $B_{ixy} = \frac{\theta_i}{n_i} \left(1 + \log \left(q_{xy} + \sum_{l \in \mathcal{I}} \frac{\theta_l \overline{K_{lxy}}}{n_l} \right) \right),$

– $C_{ijxy} = \dfrac{\theta_i \theta_j}{2n_i n_j \left(q_{xy} + \sum_{l \in \mathcal{I}} \frac{\theta_l \overline{K_{lxy}}}{n_l}\right)}$,

To compute the expectation $E(\hat{H}(X,Y))$ of the joint entropy, it should be noted that if $i \neq j$ then K_{ixy} and K_{jxy} are independent, hence $E((K_{ixy} - \overline{K_{ixy}})(K_{jxy} - \overline{K_{jxy}})) = 0$. Also note that $(K_{ixy} \colon (x,y) \in \mathcal{D})$ follows the multinomial distribution with the sample size n_i and the probabilities $\frac{R_i[x,y]}{\theta_i}$ for $(x,y) \in \mathcal{D}$, therefore

$$E\left((K_{ixy} - \overline{K_{ixy}})^2\right) = n_i \frac{R_i[x,y]}{\theta_i}\left(1 - \frac{R_i[x,y]}{\theta_i}\right) = \overline{K_{ixy}}\left(1 - \frac{\overline{K_{ixy}}}{n_i}\right).$$

We will also use $E(K_{ixy} - \overline{K_{ixy}}) = 0$, which is immediate from $\overline{K_{ixy}} = E(K_{ixy})$.

Then the expectation of $\hat{H}(X,Y)$ is given by:

$$\begin{aligned}
E(\hat{H}(X,Y)) = &-\sum_{(x,y)\in\mathcal{D}} (A_{xy} + \sum_{i \in \mathcal{I}} B_{ixy} E(K_{ixy} - \overline{K_{ixy}}) \\
&\quad + \sum_{i,j \in \mathcal{I},\, i \neq j} C_{ijxy} E\left((K_{ixy} - \overline{K_{ixy}})(K_{jxy} - \overline{K_{jxy}})\right) \\
&\quad + \sum_{i \in \mathcal{I}} C_{iixy} E\left((K_{ixy} - \overline{K_{ixy}})^2\right) + \mathcal{O}(n_i^{-2})) \\
= H(X,Y) &- \sum_{i \in \mathcal{I}} \frac{\theta_i^2}{2n_i^2} \sum_{(x,y)\in\mathcal{D}} \frac{\overline{K_{ixy}}\left(1 - \frac{\overline{K_{ixy}}}{n_i}\right)}{q_{xy} + \sum_{l \in \mathcal{I}} \frac{\theta_l \overline{K_{lxy}}}{n_l}} \\
= H(X,Y) &- \sum_{i \in \mathcal{I}} \frac{\theta_i^2}{2n_i} \sum_{(x,y)\in\mathcal{D}} \frac{D_i[x,y]\,(1 - D_i[x,y])}{P_{XY}[x,y]}
\end{aligned}$$

where we use $D_i[x,y] = \frac{R_i[x,y]}{\theta_i}$.

Next we calculate the expectation $E(\hat{H}(Y))$ of the empirical entropy of observables. For simplicity we use the following notations. Let $\mathcal{Y}^+$ be the set of observables with non-zero probabilities: $\mathcal{Y}^+ = \left\{y \in \mathcal{Y} : \sum_{x \in \mathcal{D}_y} P_{XY}[x,y] \neq 0\right\}$. Let $\mathcal{D}_x = \{y \colon (x,y) \in \mathcal{D}\}$ and $\mathcal{D}_y = \{x \colon (x,y) \in \mathcal{D}\}$. For each $i \in \mathcal{I}$ and $y \in \mathcal{Y}$ let $L_{i\cdot y} = \sum_{x \in \mathcal{D}_y} K_{ixy}$. Then the empirical entropy of observables is defined by

$$\begin{aligned}
\hat{H}(Y) &= -\sum_{y \in \mathcal{Y}^+} \left(\left(\sum_{x \in \mathcal{D}_y} \hat{P}_{XY}[x,y]\right) \log\left(\sum_{x \in \mathcal{D}_y} \hat{P}_{XY}[x,y]\right)\right) \\
&= -\sum_{y \in \mathcal{Y}^+} \left(\sum_{x \in \mathcal{D}_y} q_{xy} + \sum_{i \in \mathcal{I}} \frac{\theta_i L_{i\cdot y}}{n_i}\right) \log\left(\sum_{x \in \mathcal{D}_y} q_{xy} + \sum_{i \in \mathcal{I}} \frac{\theta_i L_{i\cdot y}}{n_i}\right).
\end{aligned}$$

Let $\overline{L_{i\cdot y}} = E(L_{i\cdot y})$. Then $\overline{L_{i\cdot y}} = \sum_{x\in\mathcal{D}_y} \overline{K_{ixy}}$. By the Taylor expansion w.r.t. $\overline{L_{i\cdot y}}$ for all $i \in \mathcal{I}$,

$$\hat{H}(Y) = -\sum_{y\in\mathcal{Y}^+} (A_{\cdot y} + \sum_{i\in\mathcal{I}} B_{i\cdot y}(L_{i\cdot y} - \overline{L_{i\cdot y}}) + \sum_{i,j\in\mathcal{I},\, i\neq j} C_{ij\cdot y}(L_{i\cdot y} - \overline{L_{i\cdot y}})(L_{j\cdot y} - \overline{L_{j\cdot y}}) + \sum_{i\in\mathcal{I}} C_{ii\cdot y}(L_{i\cdot y} - \overline{L_{i\cdot y}})^2 + \mathcal{O}(n_i^{-2}))$$

where

- $A_{\cdot y} = \left(\sum_{x\in\mathcal{D}_y} q_{xy} + \sum_{l\in\mathcal{I}} \frac{\theta_i \overline{L_{l\cdot y}}}{n_l}\right) \log\left(\sum_{x\in\mathcal{D}_y} q_{xy} + \sum_{l\in\mathcal{I}} \frac{\theta_i \overline{L_{l\cdot y}}}{n_l}\right)$,
- $B_{i\cdot y} = \frac{\theta_i}{n_i}\left(1 + \left(\sum_{x\in\mathcal{D}_y} q_{xy} + \sum_{l\in\mathcal{I}} \frac{\theta_l \overline{L_{l\cdot y}}}{n_l}\right)\right)$,
- $C_{ij\cdot y} = \frac{\theta_i\theta_j}{2n_i n_j \left(\sum_{x\in\mathcal{D}_y} q_{xy} + \sum_{l\in\mathcal{I}} \frac{\theta_l \overline{L_{l\cdot y}}}{n_l}\right)}$,

To compute the expectation $E(\hat{H}(Y))$, it should be noted that if $i \neq j$ then $L_{i\cdot y}$ and $L_{j\cdot y}$ are independent, hence $E\left((L_{i\cdot y} - \overline{L_{i\cdot y}})(L_{j\cdot y} - \overline{L_{j\cdot y}})\right) = 0$. Also note that $(L_{i\cdot y} \colon y \in \mathcal{Y}^+)$ follows the multinomial distribution with the sample size n_i and the probabilities $\frac{\sum_{x\in\mathcal{D}_y} R_i[x,y]}{\theta_i}$ for $y \in \mathcal{Y}^+$, therefore

$$\begin{aligned} E\left((L_{i\cdot y} - \overline{L_{i\cdot y}})^2\right) &= n_i \frac{\sum_{x\in\mathcal{D}_y} R_i[x,y]}{\theta_i}\left(1 - \frac{\sum_{x\in\mathcal{D}_y} R_i[x,y]}{\theta_i}\right) \\ &= \left(\sum_{x\in\mathcal{D}_y} \overline{K_{ixy}}\right)\left(1 - \frac{\sum_{x\in\mathcal{D}_y} \overline{K_{ixy}}}{n_i}\right). \end{aligned}$$

We will also use $E(L_{i\cdot y} - \overline{L_{i\cdot y}}) = 0$, which is immediate from $\overline{L_{i\cdot y}} = E(L_{i\cdot y})$.

The expectation $E(\hat{H}(Y))$ is given by:

$$\begin{aligned} E(\hat{H}(Y)) &= -\sum_{y\in\mathcal{Y}^+} (A_{\cdot y} + \sum_{i\in\mathcal{I}} B_{i\cdot y} E(L_{i\cdot y} - \overline{L_{i\cdot y}}) + \sum_{i,j\in\mathcal{I},\, i\neq j} C_{ij\cdot y} E\left((L_{i\cdot y} - \overline{L_{i\cdot y}})(L_{j\cdot y} - \overline{L_{j\cdot y}})\right) \\ &\qquad + \sum_{i\in\mathcal{I}} C_{ii\cdot y} E\left((L_{i\cdot y} - \overline{L_{i\cdot y}})^2\right) + \mathcal{O}(n_i^{-2})) \\ &= H(Y) - \sum_{i\in\mathcal{I}} \sum_{y\in\mathcal{Y}^+} C_{ii\cdot y}\left(\sum_{x\in\mathcal{D}_y} \overline{K_{ixy}}\right)\left(1 - \frac{\sum_{x\in\mathcal{D}_y} \overline{K_{ixy}}}{n_i}\right) + \mathcal{O}(n_i^{-2}). \\ &= H(Y) - \sum_{i\in\mathcal{I}} \frac{\theta_i^2}{2n_i^2} \sum_{y\in\mathcal{Y}^+} \frac{\sum_{x\in\mathcal{D}_y} \overline{K_{ixy}}\left(1 - \sum_{x\in\mathcal{D}_y} \frac{\overline{K_{ixy}}}{n_i}\right)}{\sum_{x\in\mathcal{D}_y} q_{xy} + \sum_{l\in\mathcal{I}} \frac{\theta_l \overline{K_{lxy}}}{n_l}} \\ &= H(Y) - \sum_{i\in\mathcal{I}} \frac{\theta_i^2}{2n_i} \sum_{y\in\mathcal{Y}^+} \frac{D_{Yi}[y](1 - D_{Yi}[y])}{P_Y[y]} \end{aligned}$$

where $D_{Yi}[y] = \frac{\sum_{x\in\mathcal{D}_y} \overline{K_{ixy}}}{n_i}$ for each $y \in \mathcal{Y}$.

Similarly, the expectation $E(\hat{H}(X))$ is given by:

$$E(\hat{H}(X)) = H(X) - \sum_{i\in\mathcal{I}} \frac{\theta_i^2}{2n_i} \sum_{x\in\mathcal{X}^+} \frac{D_{Xi}[x]\,(1 - D_{Xi}[x])}{P_X[x]}$$

The expectation of the mutual information $E(\hat{I}(X;Y))$ is given by:

$$\begin{aligned}
&E(\hat{I}(X;Y)) \\
&= E(\hat{H}(X)) + E(\hat{H}(Y)) - E(\hat{H}(X,Y)) \\
&= I(X;Y) + \sum_{i\in\mathcal{I}} \frac{\theta_i^2}{2n_i} \left(\sum_{(x,y)\in\mathcal{D}} \frac{D_i[x,y] - D_i[x,y]^2}{P_{XY}[x,y]} - \sum_{x\in\mathcal{X}^+} \frac{D_{Xi}[x] - D_{Xi}[x]^2}{P_X[x]} - \sum_{y\in\mathcal{Y}^+} \frac{D_{Yi}[y] - D_{Yi}[y]^2}{P_Y[y]} \right) + \mathcal{O}(n_i^{-2}).
\end{aligned}$$

Proof (Proof of Theorem 2). We derive the variance of the estimate as follows.
For any $i, i' \in \mathcal{I}$, $x, x' \in \mathcal{X}$ and $y, y' \in \mathcal{Y}$ the covariance $Cov(K_{ixy}, K_{i'x'y'})$ is given by:

$$Cov(K_{ixy}, K_{i'x'y'}) = \begin{cases} 0 & \text{if } i \neq i' \\ n_i D_i[x,y](1 - D_i[x,y]) & \text{if } i = i',\ x = x' \text{ and } y = y' \\ -n_i D_i[x,y] D_i[x',y'] & \text{otherwise.} \end{cases}$$

The covariance $Cov(L_{i\cdot y}, L_{i\cdot y'})$ depends on whether $y = y'$ or not:

$$\begin{aligned}
Cov(L_{i\cdot y}, L_{i\cdot y}) &= Cov\left(\sum_{x\in\mathcal{D}_y} K_{ixy}, \sum_{x'\in\mathcal{D}_y} K_{ix'y} \right) \\
&= \sum_{x\in\mathcal{D}_y} \sum_{x'\in\mathcal{D}_y} Cov(K_{ixy}, K_{ix'y}) \\
&= \sum_{x\in\mathcal{D}_y} n_i D_i[x,y](1 - D_i[x,y]) - \sum_{x\neq x'\in\mathcal{D}_y} n_i D_i[x,y] D_i[x',y] \\
&= n_i D_{Yi}[y](1 - D_{Yi}[y])
\end{aligned}$$

When $y \neq y'$:

$$\begin{aligned}
Cov(L_{i\cdot y}, L_{i\cdot y'}) &= Cov\left(\sum_{x\in\mathcal{D}_y} K_{ixy}, \sum_{x'\in\mathcal{D}_{y'}} K_{ix'y'} \right) \\
&= \sum_{x\in\mathcal{D}_y} \sum_{x'\in\mathcal{D}_{y'}} Cov(K_{ixy}, K_{ix'y'}) \\
&= -n_i \sum_{x\in\mathcal{D}_y} \sum_{x'\in\mathcal{D}_{y'}} D_i[x,y] D_i[x',y'] \\
&= -n_i D_{Yi}[y] D_{Yi}[y']
\end{aligned}$$

The variance of $\hat{H}(X,Y)$ is given by the following.

$$\begin{aligned}
V(\hat{H}(X,Y)) &= E\left(\hat{H}(X,Y)^2\right) - \left(E(\hat{H}(X,Y))\right)^2 \\
&= \sum_{(x,y)\in\mathcal{D}} \sum_{(x',y')\in\mathcal{D}} \sum_{i,i'\in\mathcal{I}} B_{ixy}B_{i'x'y'}\, Cov(K_{ixy}, K_{i'x'y'}) + \mathcal{O}(n_i^{-2}) \\
&= \sum_{i\in\mathcal{I}} n_i \sum_{(x,y)\in\mathcal{D}} D_i[x,y]B_{ixy}\left(B_{ixy} - \sum_{(x',y')\in\mathcal{D}} B_{ix'y'}D_i[x',y']\right) + \mathcal{O}(n_i^{-2})
\end{aligned}$$

The variance of $\hat{H}(Y)$ is given by the following.

$$\begin{aligned}
V(\hat{H}(Y)) &= E\left(\hat{H}(Y)^2\right) - \left(E(\hat{H}(Y))\right)^2 \\
&= \sum_{y\in\mathcal{Y}^+} \sum_{y'\in\mathcal{Y}^+} \sum_{i,i'\in\mathcal{I}} B_{i\cdot y}B_{i'\cdot y'}\, Cov(L_{i\cdot y}, L_{i'\cdot y'}) + \mathcal{O}(n_i^{-2}) \\
&= \sum_{i\in\mathcal{I}} n_i \sum_{y\in\mathcal{Y}^+} D_{Yi}[y]B_{i\cdot y}\left(B_{i\cdot y} - \sum_{y'\in\mathcal{Y}^+} B_{i\cdot y'}D_{Yi}[y']\right) + \mathcal{O}(n_i^{-2})
\end{aligned}$$

Similarly, the variance of $\hat{H}(X)$ is given by the following.

$$V(\hat{H}(X)) = \sum_{i\in\mathcal{I}} n_i \sum_{x\in\mathcal{X}^+} D_{Xi}[x]B_{ix\cdot}\left(B_{ix\cdot} - \sum_{x'\in\mathcal{X}^+} B_{ix'\cdot}D_{Xi}[x']\right) + \mathcal{O}(n_i^{-2})$$

The covariance between $\hat{H}(X,Y)$ and $\hat{H}(Y)$ is given by:

$$\begin{aligned}
Cov(\hat{H}(Y), \hat{H}(X,Y)) &= \sum_{i\in\mathcal{I}} \sum_{(x,y)\in\mathcal{D}} \sum_{y'\in\mathcal{Y}^+} B_{ixy}B_{i\cdot y'}\, Cov(K_{ixy}, L_{i\cdot y'}) + \mathcal{O}(n_i^{-2}) \\
&= \sum_{i\in\mathcal{I}} \sum_{(x,y)\in\mathcal{D}} \sum_{y'\in\mathcal{Y}^+} B_{ixy}B_{i\cdot y'}\, Cov\left(K_{ixy}, \sum_{x'\in\mathcal{D}_{y'}} K_{ix'y'}\right) + \mathcal{O}(n_i^{-2}) \\
&= \sum_{i\in\mathcal{I}} \sum_{(x,y)\in\mathcal{D}} B_{ixy} \sum_{(x',y')\in\mathcal{D}} B_{i\cdot y'}\, Cov(K_{ixy}, K_{ix'y'}) + \mathcal{O}(n_i^{-2}) \\
&= \sum_{i\in\mathcal{I}} n_i \sum_{(x,y)\in\mathcal{D}} D_i[x,y]B_{ixy}\left(B_{i\cdot y} - \sum_{(x',y')\in\mathcal{D}} B_{i\cdot y'}D_i[x',y']\right) + \mathcal{O}(n_i^{-2})
\end{aligned}$$

Similarly, the covariance between $\hat{H}(X,Y)$ and $\hat{H}(X)$ is given by:

$$Cov(\hat{H}(X), \hat{H}(X,Y)) = \sum_{i\in\mathcal{I}} n_i \sum_{(x,y)\in\mathcal{D}} D_i[x,y]B_{ixy}\left(B_{ix\cdot} - \sum_{(x',y')\in\mathcal{D}} B_{ix'\cdot}D_i[x',y']\right) + \mathcal{O}(n_i^{-2})$$

The covariance between $\hat{H}(X)$ and $\hat{H}(Y)$ is given by:

$$\begin{aligned}
Cov(\hat{H}(X),\hat{H}(Y)) =& \sum_{i\in\mathcal{I}}\sum_{x\in\mathcal{X}^+}\sum_{y'\in\mathcal{Y}^+} B_{ix\cdot}B_{i\cdot y'}\,Cov(L_{ix\cdot},L_{i\cdot y'})+\mathcal{O}(n_i^{-2})\\
=& \sum_{i\in\mathcal{I}}\sum_{x\in\mathcal{X}^+}\sum_{y'\in\mathcal{Y}^+} B_{ix\cdot}B_{i\cdot y'}\,Cov\Big(\sum_{y\in\mathcal{D}_x}K_{ixy},\sum_{x'\in\mathcal{D}_{y'}}K_{ix'y'}\Big)+\mathcal{O}(n_i^{-2})\\
=& \sum_{i\in\mathcal{I}}\sum_{(x,y)\in\mathcal{D}} B_{ix\cdot}\sum_{(x',y')\in\mathcal{D}} B_{i\cdot y'}\,Cov(K_{ixy},K_{ix'y'})+\mathcal{O}(n_i^{-2})\\
=& \sum_{i\in\mathcal{I}} n_i\sum_{(x,y)\in\mathcal{D}} D_i[x,y]B_{ix\cdot}\Big(B_{i\cdot y}-\sum_{(x',y')\in\mathcal{D}} B_{i\cdot y'}D_i[x',y']\Big)+\mathcal{O}(n_i^{-2})
\end{aligned}$$

Therefore the variance of the mutual information is as follows:

$$\begin{aligned}
&V(\hat{I}(X;Y))\\
&= V\Big(\hat{H}(X)+\hat{H}(Y)-\hat{H}(X,Y)\Big)\\
&= V(\hat{H}(X))+V(\hat{H}(Y))+V(\hat{H}(X,Y))\\
&\quad +2Cov(\hat{H}(X),\hat{H}(Y))-2Cov(\hat{H}(X),\hat{H}(X,Y))-2Cov(\hat{H}(Y),\hat{H}(X,Y))\\
&= \sum_{i\in\mathcal{I}} n_i\sum_{(x,y)\in\mathcal{D}} D_i[x,y]\Big(B_{ix}\Big(B_{ix}-\sum_{(x',y')\in\mathcal{D}}B_{ix'}D_i[x',y']\Big)+B_{iy}\Big(B_{iy}-\sum_{(x',y')\in\mathcal{D}}B_{iy'}D_i[x',y']\Big)\\
&\qquad +B_{ixy}\Big(B_{ixy}-\sum_{(x',y')\in\mathcal{D}}B_{ix'y'}D_i[x',y']\Big)+2B_{ix}\Big(B_{iy}-\sum_{(x',y')\in\mathcal{D}}B_{iy'}D_i[x',y']\Big)\\
&\qquad -2B_{ixy}\Big(B_{ix}-\sum_{(x',y')\in\mathcal{D}}B_{ix'}D_i[x',y']\Big)-2B_{ixy}\Big(B_{iy}-\sum_{(x',y')\in\mathcal{D}}B_{iy'}D_i[x',y']\Big)\Big)+\mathcal{O}(n_i^{-2})\\
&= \sum_{i\in\mathcal{I}}\frac{\theta_i^2}{n_i}\left(\sum_{(x,y)\in\mathcal{D}}D_i[x,y]\Big(1+\log\tfrac{P_X[x]P_Y[y]}{P_{XY}[x,y]}\Big)^2-\left(\sum_{(x,y)\in\mathcal{D}}D_i[x,y]\Big(1+\log\tfrac{P_X[x]P_Y[y]}{P_{XY}[x,y]}\Big)\right)^2\right)+\mathcal{O}(n_i^{-2})
\end{aligned}$$

A.2 Proofs for Adaptive Analysis

To prove Theorem 3 it suffices to show the following:

Theorem 6. Let $v_1, v_2, \ldots, v_m$ be m positive real numbers. Let $n, n_1, n_2, \ldots, n_m$ be $(m+1)$ positive real numbers such that $\sum_{i=1}^m n_i = n$. Then

$$\sum_{i=1}^m \frac{v_i}{n_i} \geq \frac{1}{n}\left(\sum_{i=1}^m \sqrt{v_i}\right)^2.$$

The equality holds when $n_i = \dfrac{\sqrt{v_i}n}{\sum_{j=1}^m \sqrt{v_j}}$ for all $i = 1, 2, \ldots, m$.

Proof. We prove the theorem by induction on m. When $m = 1$ the equality holds trivially. When $m = 2$ it is sufficient to prove

$$\frac{v_1}{n_1} + \frac{v_2}{n_2} \geq \frac{\left(\sqrt{v_1} + \sqrt{v_2}\right)^2}{n_1 + n_2}.$$

By $n_1, n_2 > 0$, this is equivalent to

$$(n_1 + n_2)(n_2 v_1 + n_1 v_2) \geq n_1 n_2 \left(\sqrt{v_1} + \sqrt{v_2}\right)^2.$$

We obtain this by

$$\begin{aligned}
&(n_1 + n_2)(n_2 v_1 + n_1 v_2) - n_1 n_2 \left(\sqrt{v_1} + \sqrt{v_2}\right)^2 \\
=& (n_1 + n_2) n_2 v_1 + (n_1 + n_2) n_1 v_2 - n_1 n_2 \left(v_1 + 2\sqrt{v_1 v_2} + v_2\right) \\
=& n_2^2 v_1 + n_1^2 v_2 - 2 n_1 n_2 \sqrt{v_1 v_2} \\
=& \left(n_2 \sqrt{v_1} - n_1 \sqrt{v_2}\right)^2 \\
\geq& 0.
\end{aligned}$$

Next we prove the inductive step as follows.

$$\begin{aligned}
\sum_{i=1}^{m} \frac{v_i}{n_i} &= \sum_{i=1}^{m-1} \frac{v_i}{n_i} + \frac{v_m}{n_m} \\
&\geq \frac{1}{n_1 + \ldots + n_{m-1}} \left(\sum_{i=1}^{m-1} \sqrt{v_i}\right)^2 + \frac{\sqrt{v_m}^2}{n_m} && \text{(by induction hypothesis)} \\
&\geq \frac{1}{(n_1 + \ldots + n_{m-1}) + n_m} \left(\sqrt{\left(\sum_{i=1}^{m-1} \sqrt{v_i}\right)^2} + \sqrt{v_m}\right)^2 && \text{(by induction hypothesis)} \\
&= \frac{1}{n_1 + \ldots + n_m} \left(\sum_{i=1}^{m-1} \sqrt{v_i} + \sqrt{v_m}\right)^2 \\
&= \frac{1}{n} \left(\sum_{i=1}^{m} \sqrt{v_i}\right)^2.
\end{aligned}$$

Finally, when $n_i = \frac{\sqrt{v_i} n}{\sum_{j=1}^{m} \sqrt{v_j}}$ for all $i = 1, 2, \ldots, m$, the equality holds:

$$\sum_{i=1}^{m} \frac{v_i}{n_i} = \sum_{i=1}^{m} \frac{v_i (\sum_{j=1}^{m} \sqrt{v_j})}{\sqrt{v_i} n} = \frac{1}{n} \left(\sum_{i=1}^{m} \sqrt{v_i}\right)^2.$$

A.3 Proofs for Other Results (with assuming the knowledge of the prior)

Proof (Proof of Proposition 1). Since the precise prior P_X is provided to the analyst, we have

$$E(\hat{I}(X;Y)) = E(H(X)) + E(\hat{H}(Y)) - E(\hat{H}(X,Y)).$$

By using the results on $E(\hat{H}(Y))$ and $E(\hat{H}(X,Y))$ in the proof of Theorem 1, we obtain the proposition.

Proof (Proof of Proposition 2). Since the precise prior P_X is provided to the analyst, we have

$$\begin{aligned} V(\hat{I}(X;Y)) &= V\Big(H(X) + \hat{H}(Y) - \hat{H}(X,Y)\Big) \\ &= V(H(X) + V(\hat{H}(Y)) + V(\hat{H}(X,Y)) - 2Cov(\hat{H}(Y), \hat{H}(X,Y)). \end{aligned}$$

By using the results on $V(\hat{H}(Y))$, $V(\hat{H}(X,Y))$ $Cov(\hat{H}(Y), \hat{H}(X,Y))$ in the proof of Theorem 2, we obtain Proposition 2.

Then the following proposition is straightforward from the proof of Theorem 3.

Proposition 3. *For each $i \in \mathcal{I}$ and $x \in \mathcal{X}$, let v_{ix} be the following intermediate variance of the component S_{ix}.*

$$v_{ix} = \theta_{ix}^2 \left(\sum_{y \in \mathcal{D}_x} \hat{D}_i[y|x] \left(\log \frac{\widehat{P}_Y[y]}{\widehat{P}_{XY}[x,y]} \right)^2 - \left(\sum_{y \in \mathcal{D}_x} \hat{D}_i[y|x] \left(\log \frac{\widehat{P}_Y[y]}{\widehat{P}_{XY}[x,y]} \right) \right)^2 \right).$$

Given the total sample size n, the variance of the estimated mutual information is minimized if, for all $i \in \mathcal{I}$ and $x \in \mathcal{X}$, the sample size n_i and the importance prior λ_i satisfy the following:

$$n_i \lambda_i[x] = \frac{\sqrt{v_{ix}} n}{\sum_{j=1}^{m} \sqrt{v_{jx}}} \quad .$$

A.4 Proofs for the Results on the Abstraction-Then-Sampling (with assuming the knowledge of some components)

Proof (Proof of Theorem 4). We use notations that we have introduced in the previous proofs. For each $i \in \mathcal{I}$, let $\mathcal{X}_i$ be the set of the elements of $\mathcal{X}$ that appear with non-zero probabilities in the component S_i. Recall that for each $i \in \mathcal{I}^\star$, $\pi_i[x]$ is the probability of having an input x in the component S_i.

For each $i \in \mathcal{I}^\star$ all the non-zero rows of S_i's channel matrix are the same conditional distribution; i.e., for each $x, x' \in \mathcal{X}_i$ and $y \in \mathcal{Y}_i$, $\frac{P_{XY}[x,y]}{\pi_i[x]} = \frac{P_{XY}[x',y]}{\pi_i[x']}$ when $\pi_i[x] \neq 0$ and $\pi_i[x'] \neq 0$. Therefore it is sufficient to estimate only one of the rows. We execute the component S_i with an identical input $x \in \mathcal{X}$ n_i times to record the traces. Let $K_{i \cdot y}$ be the number of traces of the component S_i that outputs y.

From these numbers of traces we define the empirical joint (full) distribution $\hat{D}_i$ of X and Y as

$$\hat{D}_i[x,y] \stackrel{\text{def}}{=} \frac{\pi_i[x] K_{i \cdot y}}{n_i}$$

Since S_i is executed with probability θ_i, the empirical sub-distribution $\hat{R}_i$ is given by

$$\hat{R}_i[x,y] \stackrel{\text{def}}{=} \theta_i \hat{D}_i[x,y] = \frac{\theta_i \pi_i[x] K_{i \cdot y}}{n_i} \quad .$$

The empirical joint probability is given by

$$\hat{P}_{XY}[x,y] = q_{xy} + \sum_{i \in \mathcal{I} \setminus \mathcal{I}^\star} \frac{\theta_i K_{ixy}}{n_i} + \sum_{i \in \mathcal{I}^\star} \frac{\theta_i \pi_i[x] K_{i\cdot y}}{n_i}.$$

Then the empirical joint entropy is defined by

$$\begin{aligned}
&\hat{H}_{\mathcal{I}^\star}(X,Y) \\
&= -\sum_{(x,y)\in\mathcal{D}} \hat{P}_{XY}[x,y] \log \hat{P}_{XY}[x,y] \\
&= -\sum_{(x,y)\in\mathcal{D}} \left(q_{xy} + \sum_{i \in \mathcal{I} \setminus \mathcal{I}^\star} \tfrac{\theta_i K_{ixy}}{n_i} + \sum_{i \in \mathcal{I}^\star} \tfrac{\theta_i \pi_i[x] K_{i\cdot y}}{n_i}\right) \log\left(q_{xy} + \sum_{i \in \mathcal{I} \setminus \mathcal{I}^\star} \tfrac{\theta_i K_{ixy}}{n_i} + \sum_{i \in \mathcal{I}^\star} \tfrac{\theta_i \pi_i[x] K_{i\cdot y}}{n_i}\right).
\end{aligned}$$

Let $\overline{K_{i\cdot y}} = E(K_{i\cdot y})$. Then $\overline{K_{i\cdot y}} = \frac{n_i D_i[x,y]}{\pi_i[x]} = \frac{n_i R_i[x,y]}{\theta_i \pi_i[x]}$. By the Taylor expansion w.r.t. $\overline{K_{ixy}}$ for all $i \in \mathcal{I} \setminus \mathcal{I}^\star$ and $\overline{K_{i\cdot y}}$ for all $i \in \mathcal{I}^\star$,

$$\begin{aligned}
\hat{H}_{\mathcal{I}^\star}(X,Y) = -\sum_{(x,y)\in\mathcal{D}} (A_{xy} &+ \sum_{i \in \mathcal{I} \setminus \mathcal{I}^\star} B_{ixy}(K_{ixy} - \overline{K_{ixy}}) + \sum_{i \in \mathcal{I}^\star} \pi_i[x] B_{ixy}(K_{i\cdot y} - \overline{K_{i\cdot y}}) \\
&+ \sum_{i,j \in \mathcal{I} \setminus \mathcal{I}^\star,\, i \neq j} C_{ijxy}(K_{ixy} - \overline{K_{ixy}})(K_{jxy} - \overline{K_{jxy}}) \\
&+ \sum_{i \in \mathcal{I} \setminus \mathcal{I}^\star,\, j \in \mathcal{I}^\star} \pi_j[x] C_{ijxy}(K_{ixy} - \overline{K_{ixy}})(K_{j\cdot y} - \overline{K_{j\cdot y}}) \\
&+ \sum_{i,j \in \mathcal{I}^\star,\, i \neq j} \pi_i[x]\pi_j[x] C_{ijxy}(K_{i\cdot y} - \overline{K_{i\cdot y}})(K_{j\cdot y} - \overline{K_{j\cdot y}}) \\
&+ \sum_{i \in \mathcal{I} \setminus \mathcal{I}^\star} C_{iixy}(K_{ixy} - \overline{K_{ixy}})^2 + \sum_{i \in \mathcal{I}^\star} \pi_i[x]^2 C_{iixy}(K_{i\cdot y} - \overline{K_{i\cdot y}})^2 + \mathcal{O}(n_i^{-2}))
\end{aligned}$$

where

- $A_{xy} = \left(q_{xy} + \sum_{l \in \mathcal{I} \setminus \mathcal{I}^\star} \frac{\theta_l \overline{K_{lxy}}}{n_l} + \sum_{l \in \mathcal{I}^\star} \frac{\theta_l \pi_l[x] \overline{K_{l\cdot y}}}{n_l}\right) \log\left(q_{xy} + \sum_{l \in \mathcal{I} \setminus \mathcal{I}^\star} \frac{\theta_l \overline{K_{lxy}}}{n_l} + \sum_{l \in \mathcal{I}^\star} \frac{\theta_l \pi_l[x] \overline{K_{l\cdot y}}}{n_l}\right)$,
- $B_{ixy} = \frac{\theta_i}{n_i}\left(1 + \log\left(q_{xy} + \sum_{l \in \mathcal{I} \setminus \mathcal{I}^\star} \frac{\theta_l \overline{K_{lxy}}}{n_l} + \sum_{l \in \mathcal{I}^\star} \frac{\theta_l \pi_l[x] \overline{K_{l\cdot y}}}{n_l}\right)\right)$,
- $C_{ijxy} = \dfrac{\theta_i \theta_j}{2 n_i n_j \left(q_{xy} + \sum_{l \in \mathcal{I} \setminus \mathcal{I}^\star} \frac{\theta_l \overline{K_{lxy}}}{n_l} + \sum_{l \in \mathcal{I}^\star} \frac{\theta_l \pi_l[x] \overline{K_{l\cdot y}}}{n_l}\right)}$,

To compute the expectation $E(\hat{H}_{\mathcal{I}^\star}(X,Y))$ of the joint entropy, it should be noted that if $i \neq j$ then K_{ixy} and K_{jxy} are independent, hence $E((K_{ixy} - \overline{K_{ixy}})(K_{jxy} - \overline{K_{jxy}})) = 0$. Similarly, $E((K_{ixy} - \overline{K_{ixy}})(K_{j\cdot y} - \overline{K_{j\cdot y}})) = 0$ and $E((K_{i\cdot y} - \overline{K_{i\cdot y}})(K_{j\cdot y} - \overline{K_{j\cdot y}})) = 0$. Also note that for each $i \in \mathcal{I}^\star$, $(K_{i\cdot y} \colon y \in \mathcal{Y}^+)$ follows the multinomial distribution with the sample size n_i and the probabilities $\frac{R_i[x,y]}{\theta_i \pi_i[x]}$ for $(x,y) \in \mathcal{D}$, therefore

$$E\left((K_{i\cdot y} - \overline{K_{i\cdot y}})^2\right) = n_i \frac{R_i[x,y]}{\theta_i \pi_i[x]} \left(1 - \frac{R_i[x,y]}{\theta_i \pi_i[x]}\right) = \overline{K_{i\cdot y}} \left(1 - \frac{\overline{K_{i\cdot y}}}{n_i}\right).$$

We will also use $E(K_{i\cdot y} - \overline{K_{i\cdot y}}) = 0$, which is immediate from $\overline{K_{i\cdot y}} = E(K_{i\cdot y})$.

Then the expectation of $\hat{H}_{\mathcal{I}^\star}(X, Y)$ is given by:

$$
\begin{aligned}
&E(\hat{H}_{\mathcal{I}^\star}(X,Y)) \\
&= -\sum_{(x,y)\in\mathcal{D}}(A_{xy} + \sum_{i\in\mathcal{I}\setminus\mathcal{I}^\star} B_{ixy}E(K_{ixy} - \overline{K_{ixy}}) + \sum_{i\in\mathcal{I}^\star}\pi_i[x]B_{ixy}E(K_{i\cdot y} - \overline{K_{i\cdot y}}) \\
&\quad + \sum_{i,j\in\mathcal{I}\setminus\mathcal{I}^\star,\, i\neq j} C_{ijxy}E\left((K_{ixy} - \overline{K_{ixy}})(K_{jxy} - \overline{K_{jxy}})\right) \\
&\quad + \sum_{i\in\mathcal{I}\setminus\mathcal{I}^\star,\, j\in\mathcal{I}^\star} \pi_j[x]C_{ijxy}E\left((K_{ixy} - \overline{K_{ixy}})(K_{j\cdot y} - \overline{K_{j\cdot y}})\right) \\
&\quad + \sum_{i,j\in\mathcal{I}^\star,\, i\neq j} \pi_i[x]\pi_j[x]C_{ijxy}E\left((K_{i\cdot y} - \overline{K_{i\cdot y}})(K_{j\cdot y} - \overline{K_{j\cdot y}})\right) \\
&\quad + \sum_{i\in\mathcal{I}\setminus\mathcal{I}^\star} C_{iixy}E\left((K_{ixy} - \overline{K_{ixy}})^2\right) + \sum_{i\in\mathcal{I}^\star}\pi_i[x]^2C_{iixy}E\left((K_{i\cdot y} - \overline{K_{i\cdot y}})^2\right) + \mathcal{O}(n_i^{-2})) \\
&= H(X,Y) - \sum_{i\in\mathcal{I}\setminus\mathcal{I}^\star}\frac{\theta_i^2}{2n_i^2}\sum_{(x,y)\in\mathcal{D}}\frac{\overline{K_{ixy}}\left(1-\frac{\overline{K_{ixy}}}{n_i}\right)}{P_{XY}[x,y]} - \sum_{i\in\mathcal{I}^\star}\frac{\theta_i^2\pi_i[x]^2}{2n_i^2}\sum_{(x,y)\in\mathcal{D}}\frac{\overline{K_{i\cdot y}}\left(1-\frac{\overline{K_{i\cdot y}}}{n_i}\right)}{P_{XY}[x,y]} \\
&= H(X,Y) - \sum_{i\in\mathcal{I}\setminus\mathcal{I}^\star}\frac{\theta_i^2}{2n_i}\sum_{(x,y)\in\mathcal{D}}\frac{D_i[x,y](1-D_i[x,y])}{P_{XY}[x,y]} - \sum_{i\in\mathcal{I}^\star}\frac{\theta_i^2}{2n_i}\sum_{(x,y)\in\mathcal{D}}\frac{D_i[x,y](\pi_i[x]-D_i[x,y])}{P_{XY}[x,y]}
\end{aligned}
$$

where we use $\overline{K_{i\cdot y}} = \frac{n_i D_i[x,y]}{\pi_i[x]}$.

Next we calculate the expectation $E(\hat{H}_{\mathcal{I}^\star}(Y))$ of the empirical entropy of observables. Recall that $\mathcal{Y}^+$ is the set of observables with non-zero probabilities, $\mathcal{D}$ is the set of pairs of secrets and observables with non-zero probabilities, $\mathcal{D}_x = \{y\colon (x,y)\in\mathcal{D}\}$ and $\mathcal{D}_y = \{x\colon (x,y)\in\mathcal{D}\}$. For each $i\in\mathcal{I}\setminus\mathcal{I}^\star$ and $y\in\mathcal{Y}$ let $L_{i\cdot y} = \sum_{x\in\mathcal{D}_y} K_{ixy}$. Then the empirical entropy of observables is defined by

$$\hat{H}_{\mathcal{I}^\star}(Y) = -\sum_{y\in\mathcal{Y}^+}\Bigg(\sum_{x\in\mathcal{D}_y} q_{xy} + \sum_{i\in\mathcal{I}\setminus\mathcal{I}^\star}\frac{\theta_i L_{i\cdot y}}{n_i} + \sum_{i\in\mathcal{I}^\star}\frac{\theta_i K_{i\cdot y}}{n_i}\Bigg)\log\Bigg(\sum_{x\in\mathcal{D}_y} q_{xy} + \sum_{i\in\mathcal{I}\setminus\mathcal{I}^\star}\frac{\theta_i L_{i\cdot y}}{n_i} + \sum_{i\in\mathcal{I}^\star}\frac{\theta_i K_{i\cdot y}}{n_i}\Bigg).$$

For $i\in\mathcal{I}\setminus\mathcal{I}^\star$, let $\overline{L_{i\cdot y}} = E(L_{i\cdot y})$. Then $\overline{L_{i\cdot y}} = \sum_{x\in\mathcal{D}_y}\overline{K_{ixy}}$.

- $A_{\cdot y} = \left(\sum_{x\in\mathcal{D}_y} q_{xy} + \sum_{l\in\mathcal{I}\setminus\mathcal{I}^\star}\frac{\theta_i\overline{L_{l\cdot y}}}{n_l} + \sum_{l\in\mathcal{I}^\star}\frac{\theta_l\overline{K_{l\cdot y}}}{n_l}\right)\log\left(\sum_{x\in\mathcal{D}_y} q_{xy} + \sum_{l\in\mathcal{I}\setminus\mathcal{I}^\star}\frac{\theta_i\overline{L_{l\cdot y}}}{n_l} + \sum_{l\in\mathcal{I}^\star}\frac{\theta_l\overline{K_{l\cdot y}}}{n_l}\right)$,
- $B_{i\cdot y} = \frac{\theta_i}{n_i}\left(1 + \log\left(\sum_{x\in\mathcal{D}_y} q_{xy} + \sum_{l\in\mathcal{I}\setminus\mathcal{I}^\star}\frac{\theta_l\overline{L_{l\cdot y}}}{n_l} + \sum_{l\in\mathcal{I}^\star}\frac{\theta_l\overline{K_{l\cdot y}}}{n_l}\right)\right)$,
- $C_{ij\cdot y} = \dfrac{\theta_i\theta_j}{2n_in_j\left(\sum_{x\in\mathcal{D}_y} q_{xy} + \sum_{l\in\mathcal{I}\setminus\mathcal{I}^\star}\frac{\theta_l\overline{L_{l\cdot y}}}{n_l} + \sum_{l\in\mathcal{I}^\star}\frac{\theta_l\overline{K_{l\cdot y}}}{n_l}\right)}$,

To compute the expectation $E(\hat{H}_{\mathcal{I}^\star}(Y))$, it should be noted that if $i \neq j$ then $L_{i\cdot y}$ and $L_{j\cdot y}$ are independent, hence $E\left((L_{i\cdot y} - \overline{L_{i\cdot y}})(L_{j\cdot y} - \overline{L_{j\cdot y}})\right) = 0$. Similarly, $E\left((L_{i\cdot y} - \overline{L_{i\cdot y}})(K_{j\cdot y} - \overline{K_{j\cdot y}})\right) = 0$ and $E\left((K_{i\cdot y} - \overline{K_{i\cdot y}})(K_{j\cdot y} - \overline{K_{j\cdot y}})\right) = 0$.

By the Taylor expansion w.r.t. $\overline{L_{i\cdot y}}$ for all $i \in \mathcal{I} \setminus \mathcal{I}^\star$ and $\overline{K_{i\cdot y}}$ for all $i \in \mathcal{I}^\star$, the expectation $E(\hat{H}_{\mathcal{I}^\star}(Y))$ is given by:

$$\begin{aligned} E(\hat{H}_{\mathcal{I}^\star}(Y)) &= H(Y) - \sum_{y\in\mathcal{Y}^+}\left(\sum_{i\in\mathcal{I}\setminus\mathcal{I}^\star} C_{ii\cdot y} E\left((L_{i\cdot y} - \overline{L_{i\cdot y}})^2\right) + \sum_{i\in\mathcal{I}^\star} C_{ii\cdot y} E\left((K_{i\cdot y} - \overline{K_{i\cdot y}})^2\right)\right) + \mathcal{O}(n_i^{-2}) \\ &= H(Y) - \sum_{i\in\mathcal{I}} \frac{\theta_i^2}{2n_i} \sum_{y\in\mathcal{Y}^+} \frac{D_{Yi}[y]\left(1 - D_{Yi}[y]\right)}{P_Y[y]} + \mathcal{O}(n_i^{-2}) \end{aligned}$$

where for each $i \in \mathcal{I}^\star$, $D_{Yi}[y] = \frac{\overline{K_{i\cdot y}}}{n_i} = \frac{D_i[x,y]}{\pi_i[x]}$ for all $x \in \mathcal{X}$.

The expectation $E(\hat{H}_{\mathcal{I}^\star}(X))$ is given by:

$$E(\hat{H}_{\mathcal{I}^\star}(X)) = H(X) - \sum_{i\in\mathcal{I}\setminus\mathcal{I}^\star} \frac{\theta_i^2}{2n_i} \sum_{x\in\mathcal{X}^+} \frac{D_{Xi}[x]\left(1 - D_{Xi}[x]\right)}{P_X[x]}$$

The expectation of the mutual information $E(\hat{I}_{\mathcal{I}^\star}(X;Y))$ is given by:

$$\begin{aligned} &E(\hat{I}_{\mathcal{I}^\star}(X;Y)) \\ &= E(\hat{H}_{\mathcal{I}^\star}(X)) + E(\hat{H}_{\mathcal{I}^\star}(Y)) - E(\hat{H}_{\mathcal{I}^\star}(X,Y)) \\ &= I(X;Y) + \sum_{i\in\mathcal{I}\setminus\mathcal{I}^\star} \frac{\theta_i^2}{2n_i}\left(\sum_{(x,y)\in\mathcal{D}} \frac{D_i[x,y] - D_i[x,y]^2}{P_{XY}[x,y]} - \sum_{x\in\mathcal{X}^+} \frac{D_{Xi}[x] - D_{Xi}[x]^2}{P_X[x]} - \sum_{y\in\mathcal{Y}^+} \frac{D_{Yi}[y] - D_{Yi}[y]^2}{P_Y[y]}\right) \\ &\quad + \sum_{i\in\mathcal{I}^\star} \frac{\theta_i^2}{2n_i}\left(\sum_{x\in\mathcal{D}_x} \frac{D_i[x,y]\pi_i[x] - D_i[x,y]^2}{P_{XY}[x,y]} - \sum_{y\in\mathcal{Y}^+} \frac{D_{Yi}[y] - D_{Yi}[y]^2}{P_Y[y]}\right) + \mathcal{O}(n_i^{-2}). \end{aligned}$$

Proof (Proof of Theorem 5). We derive the variance of the estimate as follows. We will use the covariances $Cov(K_{ixy}, K_{i'x'y'})$ and $Cov(L_{i\cdot y}, L_{i'\cdot y'})$ shown in the proof of Theorem 2.

For each $i, i' \in \mathcal{I}^\star$ and $y, y' \in \mathcal{Y}$ the covariance $Cov(K_{i\cdot y}, K_{i'\cdot y'})$ is as follows:

$$Cov(K_{i\cdot y}, K_{i'\cdot y'}) = \begin{cases} 0 & \text{if } i \neq i' \\ \frac{n_i D_i[x,y](\pi_i[x] - D_i[x,y])}{\pi_i[x]^2} = n_i D_{Yi}[y](1 - D_{Yi}[y]) & \text{if } i = i' \text{ and } y = y' \\ -\frac{n_i D_i[x,y] D_i[x,y']}{\pi_i[x]^2} = -n_i D_{Yi}[y] D_{Yi}[y'] & \text{otherwise.} \end{cases}$$

where x is an arbitraly element of $\mathcal{X}_i$

The variance of $\hat{H}_{\mathcal{I}^\star}(X,Y)$ is given by the following.

$$\begin{aligned}
V(\hat{H}_{\mathcal{I}^\star}(X,Y)) &= E\left(\hat{H}_{\mathcal{I}^\star}(X,Y)^2\right) - \left(E(\hat{H}_{\mathcal{I}^\star}(X,Y))\right)^2 \\
&= \sum_{i,i'\in\mathcal{I}\setminus\mathcal{I}^\star} \sum_{(x,y)\in\mathcal{D}} \sum_{(x',y')\in\mathcal{D}} B_{ixy}B_{i'x'y'}\, Cov(K_{ixy}, K_{i'x'y'}) \\
&\quad + \sum_{i,i'\in\mathcal{I}^\star} \sum_{(x,y)\in\mathcal{D}} \sum_{(x',y')\in\mathcal{D}} \pi_i[x]B_{ixy}\pi_{i'}[x']B_{i'x'y'}\, Cov(K_{i\cdot y}, K_{i'\cdot y'}) + \mathcal{O}(n_i^{-2}) \\
&= \sum_{i\in\mathcal{I}\setminus\mathcal{I}^\star} n_i \sum_{(x,y)\in\mathcal{D}} D_i[x,y]B_{ixy}\left(B_{ixy} - \sum_{(x',y')\in\mathcal{D}} B_{ix'y'}D_i[x',y']\right) \\
&\quad + \sum_{i\in\mathcal{I}^\star} n_i \sum_{(x,y)\in\mathcal{D}} \sum_{x'\in\mathcal{X}^+} D_{Yi}[y]\pi_i[x]\pi_i[x']B_{ixy}\left(B_{ix'y} - \sum_{y'\in\mathcal{D}_{x'}} B_{ix'y'}D_{Yi}[y']\right) + \mathcal{O}(n_i^{-2})
\end{aligned}$$

The variances of $\hat{H}_{\mathcal{I}^\star}(Y)$ and $\hat{H}_{\mathcal{I}^\star}(X)$ are given by the following.

$$\begin{aligned}
V(\hat{H}_{\mathcal{I}^\star}(Y)) &= \sum_{y,y'\in\mathcal{Y}^+}\left(\sum_{i,i'\in\mathcal{I}\setminus\mathcal{I}^\star} B_{i\cdot y}B_{i'\cdot y'}\, Cov(L_{i\cdot y}, L_{i'\cdot y'}) + \sum_{i,i'\in\mathcal{I}^\star} B_{i\cdot y}B_{i'\cdot y'}\, Cov(K_{i\cdot y}, K_{i'\cdot y'})\right) + \mathcal{O}(n_i^{-2}) \\
&= \sum_{i\in\mathcal{I}} n_i \sum_{y\in\mathcal{Y}^+} D_{Yi}[y]B_{i\cdot y}\left(B_{i\cdot y} - \sum_{y'\in\mathcal{Y}^+} B_{i\cdot y'}D_{Yi}[y']\right) + \mathcal{O}(n_i^{-2}) \\
V(\hat{H}_{\mathcal{I}^\star}(X)) &= \sum_{i\in\mathcal{I}\setminus\mathcal{I}^\star} n_i \sum_{x\in\mathcal{X}^+} D_{Xi}[x]B_{ix\cdot}\left(B_{ix\cdot} - \sum_{x'\in\mathcal{X}^+} B_{ix'\cdot}D_{Xi}[x']\right) + \mathcal{O}(n_i^{-2})
\end{aligned}$$

The covariance between $\hat{H}_{\mathcal{I}^\star}(X,Y)$ and $\hat{H}_{\mathcal{I}^\star}(Y)$ is given by:

$$\begin{aligned}
Cov(\hat{H}_{\mathcal{I}^\star}(Y), \hat{H}_{\mathcal{I}^\star}(X,Y)) &= \sum_{i\in\mathcal{I}\setminus\mathcal{I}^\star} \sum_{(x,y)\in\mathcal{D}} \sum_{y'\in\mathcal{Y}^+} B_{ixy}B_{i\cdot y'}\, Cov(K_{ixy}, L_{i\cdot y'}) \\
&\quad + \sum_{i\in\mathcal{I}^\star} \sum_{(x,y)\in\mathcal{D}} \sum_{y'\in\mathcal{Y}^+} \pi_i[x]B_{ixy}B_{i\cdot y'}\, Cov(K_{i\cdot y}, K_{i\cdot y'}) + \mathcal{O}(n_i^{-2}) \\
&= \sum_{i\in\mathcal{I}\setminus\mathcal{I}^\star} n_i \sum_{(x,y)\in\mathcal{D}} D_i[x,y]B_{ixy}\left(B_{i\cdot y} - \sum_{(x',y')\in\mathcal{D}} B_{i\cdot y'}D_i[x',y']\right) \\
&\quad + \sum_{i\in\mathcal{I}^\star} n_i \sum_{(x,y)\in\mathcal{D}} \pi_i[x]D_{Yi}[y]B_{ixy}\left(B_{i\cdot y} - \sum_{y'\in\mathcal{Y}^+} B_{i\cdot y'}D_{Yi}[y']\right) + \mathcal{O}(n_i^{-2})
\end{aligned}$$

Similarly, the covariance between $\hat{H}_{\mathcal{I}^\star}(X,Y)$ and $\hat{H}_{\mathcal{I}^\star}(X)$ is given by:

$$Cov(\hat{H}_{\mathcal{I}^\star}(X), \hat{H}_{\mathcal{I}^\star}(X,Y)) = \sum_{i\in\mathcal{I}\setminus\mathcal{I}^\star} n_i \sum_{(x,y)\in\mathcal{D}} D_i[x,y]B_{ixy}\left(B_{ix\cdot} - \sum_{(x',y')\in\mathcal{D}} B_{ix'\cdot}D_i[x',y']\right) + \mathcal{O}(n_i^{-2})$$

The covariance between $\hat{H}_{\mathcal{I}^\star}(X)$ and $\hat{H}_{\mathcal{I}^\star}(Y)$ is given by:

$$Cov(\hat{H}_{\mathcal{I}^\star}(X), \hat{H}_{\mathcal{I}^\star}(Y)) = \sum_{i\in\mathcal{I}\setminus\mathcal{I}^\star} n_i \sum_{(x,y)\in\mathcal{D}} D_i[x,y]B_{ix\cdot}\left(B_{i\cdot y} - \sum_{(x',y')\in\mathcal{D}} B_{i\cdot y'}D_i[x',y']\right) + \mathcal{O}(n_i^{-2})$$

Therefore the variance of the mutual information is as follows:

$$
\begin{aligned}
&V(\hat{I}_{\mathcal{I}^\star}(X;Y)) \\
&= V\Big(\hat{H}_{\mathcal{I}^\star}(X) + \hat{H}_{\mathcal{I}^\star}(Y) - \hat{H}_{\mathcal{I}^\star}(X,Y)\Big) \\
&= V(\hat{H}_{\mathcal{I}^\star}(X)) + V(\hat{H}_{\mathcal{I}^\star}(Y)) + V(\hat{H}_{\mathcal{I}^\star}(X,Y)) + 2Cov(\hat{H}_{\mathcal{I}^\star}(X), \hat{H}_{\mathcal{I}^\star}(Y)) \\
&\quad - 2Cov(\hat{H}_{\mathcal{I}^\star}(X), \hat{H}_{\mathcal{I}^\star}(X,Y)) - 2Cov(\hat{H}_{\mathcal{I}^\star}(Y), \hat{H}_{\mathcal{I}^\star}(X,Y)) \\
&= \sum_{i\in\mathcal{I}\setminus\mathcal{I}^\star} n_i \sum_{(x,y)\in\mathcal{D}} D_i[x,y]\Big(B_{ix}\Big(B_{ix} - \sum_{(x',y')\in\mathcal{D}} B_{ix'} D_i[x',y']\Big) + B_{iy}\Big(B_{iy} - \sum_{(x',y')\in\mathcal{D}} B_{iy'} D_i[x',y']\Big) \\
&\qquad\qquad + B_{ixy}\Big(B_{ixy} - \sum_{(x',y')\in\mathcal{D}} B_{ix'y'} D_i[x',y']\Big) + 2B_{ix}\Big(B_{iy} - \sum_{(x',y')\in\mathcal{D}} B_{iy'} D_i[x',y']\Big) \\
&\qquad\qquad - 2B_{ixy}\Big(B_{ix} - \sum_{(x',y')\in\mathcal{D}} B_{ix'} D_i[x',y']\Big) - 2B_{ixy}\Big(B_{iy} - \sum_{(x',y')\in\mathcal{D}} B_{iy'} D_i[x',y']\Big)\Big)
\end{aligned}
$$

$$
\begin{aligned}
&+ \sum_{i\in\mathcal{I}^\star} n_i \sum_{y\in\mathcal{Y}^+} D_{Yi}[y]\Big(B_{i\cdot y}\Big(B_{i\cdot y} - \sum_{y'\in\mathcal{Y}^+} B_{i\cdot y'} D_{Yi}[y']\Big) \\
&\qquad\qquad + \sum_{x\in\mathcal{D}_y} \pi_i[x] B_{ixy} \sum_{x'\in\mathcal{X}^+} \pi_i[x']\Big(B_{ix'y} - \sum_{y'\in\mathcal{D}_{x'}} B_{ix'y'} D_{Yi}[y']\Big)\Big) \\
&\qquad\qquad - 2\sum_{x\in\mathcal{D}_y} \pi_i[x] B_{ixy}\Big(B_{i\cdot y} - \sum_{y'\in\mathcal{Y}^+} B_{i\cdot y'} D_{Yi}[y']\Big) + \mathcal{O}(n_i^{-2}) \\
&= \sum_{i\in\mathcal{I}\setminus\mathcal{I}^\star} \frac{\theta_i^2}{n_i}\Bigg(\sum_{(x,y)\in\mathcal{D}} D_i[x,y]\Big(1 + \log\tfrac{P_X[x]P_Y[y]}{P_{XY}[x,y]}\Big)^2 - \Big(\sum_{(x,y)\in\mathcal{D}} D_i[x,y]\Big(1 + \log\tfrac{P_X[x]P_Y[y]}{P_{XY}[x,y]}\Big)\Big)^2\Bigg) \\
&+ \sum_{i\in\mathcal{I}^\star} \frac{\theta_i^2}{n_i}\Bigg(\sum_{y\in\mathcal{Y}^+} D_{Yi}[y]\Big(\log P_Y[y] - \sum_{x\in\mathcal{X}} \pi_i[x]\log P_{XY}[x,y]\Big)^2 \\
&\qquad\qquad - \Big(\sum_{y\in\mathcal{Y}^+} D_{Yi}[y]\Big(\log P_Y[y] - \sum_{x\in\mathcal{X}} \pi_i[x]\log P_{XY}[x,y]\Big)\Big)^2\Bigg) + \mathcal{O}(n_i^{-2})
\end{aligned}
$$

B Estimation of Other Measures and their Confidence Intervals

In this section we show that the hybrid statistical estimation method can be used to estimate the Shannon entropy and conditional Shannon entropy.

We recall the definitions of these measures as follows. Given a prior P_X on input X, the *prior uncertainty* (before observing the system's output Y) is defined as

$$H(X) = -\sum_{x\in\mathcal{X}} P_X[x]\log_2 P_X[x]$$

while the *posterior uncertainty* (after observing the system's output Y) is defined as

$$H(X|Y) = -\sum_{y\in\mathcal{Y}^+} P_Y[y] \sum_{x\in\mathcal{X}} P_{X|Y}[x|y] \log_2 P_{X|Y}[x|y],$$

where P_Y is the probability distribution on the output Y, $\mathcal{Y}^+$ is the set of outputs in $\mathcal{Y}$ with non-zero probabilities, and $P_{X|Y}$ is the conditional probability distribution of X given Y:

$$P_Y[y] = \sum_{x'\in\mathcal{X}} P_{XY}[x', y] \qquad P_{X|Y}[x|y] = \frac{P_{XY}[x, y]}{P_Y[y]} \quad \text{if } P_Y[y] \neq 0.$$

$H(X|Y)$ is also called the *conditional entropy* of X given Y.

B.1 Estimation of Shannon Entropy

The new method can also be used to estimate the Shannon entropy $H(X)$ of a random variable X in a probabilistic system. For each $i \in \mathcal{I}$ let D_{Xi} be the sub-distribution of X for the component S_i. Then the expectation and variance of the estimate are obtained in the same way as in the previous sections.

Proposition 4. *The expectation $E(\hat{H}(X))$ of the Shannon entropy is given by*

$$H(X) - \sum_{i\in\mathcal{I}} \frac{\theta_i^2}{2n_i} \sum_{x\in\mathcal{X}^+} \frac{D_{Xi}[x]\,(1 - D_{Xi}[x])}{P_X[x]} + \mathcal{O}(n_i^{-2})\,.$$

Proposition 5. *The variance $V(\hat{H}(X))$ of the Shannon entropy is given by*

$$\sum_{i\in\mathcal{I}} \frac{\theta_i^2}{n_i} \left(\sum_{x\in\mathcal{X}^+} D_{Xi}[x] \Big(1 + \log P_X[x]\Big)^2 - \left(\sum_{x\in\mathcal{X}^+} D_{Xi}[x] \Big(1 + \log P_X[x]\Big) \right)^2 \right) + \mathcal{O}(n_i^{-2})\,.$$

Proof. By the proof of Theorem 1 and 2 we obtain $E(\hat{H}(X))$ and $V(\hat{H}(X))$.

From these we obtain the bias and confidence interval of the Shannon entropy estimates.

Note that these propositions can respectively be derived from Theorems 1 and 2, because the Shannon entropy coincides with the mutual information of a random variable with itself: $\hat{H}(X) = \hat{I}(X; X)$. Therefore we can apply Theorem 3 and adaptively obtain the optimal sample sizes n_i.

B.2 Estimation of Conditional Entropy

The new method can also estimate the conditional Shannon entropy $H(X|Y)$ of a random variable Y given a random variable X in a probabilistic system. Intuitively, $H(X|Y)$ represents the uncertainty of a secret X after observing an output Y of the system. The expectation and variance of the conditional entropy are obtained from those of the mutual information in the case where the analyst knows the prior.

Secret \ Observable	0	1	2	3	4	5	6	7	8	9
0	0.2046	0.1102	0.0315	0.0529	0.1899	0.0064	0.0791	0.1367	0.0386	0.1501
1	0.0852	0.0539	0.1342	0.0567	0.1014	0.1254	0.0554	0.1115	0.0919	0.1844
2	0.1702	0.0542	0.0735	0.0914	0.0639	0.1322	0.1119	0.0512	0.1172	0.1343
3	0.0271	0.1915	0.0764	0.1099	0.0982	0.0761	0.0843	0.1364	0.0885	0.1116
4	0.0957	0.1977	0.0266	0.0741	0.1496	0.2177	0.0610	0.0617	0.0841	0.0318
5	0.0861	0.1275	0.1565	0.1193	0.1321	0.1716	0.0136	0.0984	0.0183	0.0766
6	0.0173	0.1481	0.1371	0.1037	0.1834	0.0271	0.1289	0.1690	0.0036	0.0818
7	0.0329	0.0825	0.0333	0.1622	0.1530	0.1378	0.0561	0.1479	0.0212	0.1731
8	0.1513	0.0435	0.0527	0.2022	0.0189	0.2159	0.0718	0.0063	0.1307	0.1067
9	0.0488	0.1576	0.1871	0.1117	0.1453	0.0349	0.0549	0.1766	0.0271	0.056

Fig. 8: Channel matrix for the experiments in Section 6.1.

Proposition 6. *The expectation $E(\hat{H}(X|Y))$ of the conditional Shannon entropy is given by $H(X) - E(\hat{I}(X;Y))$ where $E(\hat{I}(X;Y))$ is the expectation of the mutual information in the case where the analyst knows the prior (shown in Proposition 1).*

Proof. By $\hat{H}(X|Y) = H(X) - \hat{I}(X;Y)$, we obtain $E(\hat{H}(X|Y)) = H(X) - E(\hat{I}(X;Y))$. Therefore the proposition follows.

Proposition 7. *The variance $V(\hat{H}(X|Y))$ of the conditional Shannon entropy coincides with the variance $V(\hat{I}(X;Y))$ of the mutual information in the case where the analyst knows the prior (shown in Proposition 2).*

Proof. By $\hat{H}(X|Y) = H(X) - \hat{I}(X;Y)$, we obtain $V(\hat{H}(X|Y)) = V(\hat{I}(X;Y))$. Therefore the proposition follows.

C Details of Evaluation

In this section we describe more details of the discussion on the tradeoff between the cost and quality of the estimation in Section 6.1.

In Fig. 3 we showed the sampling distribution of the mutual information estimate of the joint distribution in Fig. 1 in Section 1. The graph is obtained from 1000 samples each of which is generated by combining trace analysis on a component and statistical analysis on 3 components (using 5000 randomly generated traces).

In Fig. 4a we illustrated the relationships between the size of the confidence interval and the sample size in the statistical analysis. We used an example with the randomly generated 10×10 channel matrix presented in Fig. 8 and the uniform prior. The graph shows the frequency (on the y axis) of the corrected mutual information estimates (on the x axis) that are obtained by estimating the mutual information value 1000, 5000 and 10000 times. When the sample size is k times larger then the confidence interval is $\sqrt{k}$ times narrower.

In Fig. 4b we illustrated the relationships between the size of the confidence interval and the amount of precise analysis. The graph shows the frequency (on the y axis) of the corrected mutual information estimates (on the x axis) that are obtained by estimating the mutual information value 1000 times when statistical analysis is applied to a 10×2, 10×5 and 10×10 sub-matrix of the full 10×10 matrix. Using statistical analysis only on a smaller component (10×2 sub-matrix) yields a smaller confidence interval than using it on the whole system (10×10 matrix). More generally, if we perform precise analysis on larger components, then we have a smaller confidence interval.

D On the Division into Components of the Shifting Window Benchmark

In this section we briefly discuss how the Shifting Window benchmark presented in Fig. 7 can be divided into components using the procedure in Fig. 2.

As discussed, the division into components follows the conditional branching, and since the benchmark has only one conditional statement in Line 7 we can divide into two components. We will call component S the one corresponding to the `then` branch of the conditional (i.e. Line 8) and component T the one corresponding to the `else` branch of the conditional (i.e. Line 10). The probability of each component corresponds to that of the conditional guard being true or false, and can be computed from the prior distribution on the secret.

Now for each component we need to determine which type of analysis to use. For both components it is easy to see that the output does not depend on the value of the secret, so we mark them for the abstraction-then-sampling technique of Section 4.3, meaning that their behavior can be analyzed once and the results used for all possible values of the secret. (Note that the branching depends on the secret value, which indirectly causes an information leakage in the entire system.)

We compare the estimated costs of precise and statistical analyses on the component S as follows. Since S is not deterministic, we cannot immediately mark it for precise analysis. As explained in Section 5 the choice now depends on the size of the output of the component $\#\mathcal{Y}_S$ compared to the size of the internal randomness of the component $\#\mathcal{Z}_S$. We obtain that $\#\mathcal{Y}_S = N$ and $\#\mathcal{Z}_S = W(N - W)$. Since we set $N = 2W$ and $N \geq 20$, we obtain $\#\mathcal{Z}_S = W^2 \gg 2W = \#\mathcal{Y}_S$. Consequently, precise trace analysis is more expensive than statistical analysis on the component S, thus we mark S for statistical analysis.

Component T has N possible outputs and N possible internal randomness values, so we mark it for precise analysis.

Tool	Result	Error
QUAIL	0.50327	0
LeakWatch	0.36245	0.14082
Hybrid	0.50325	0.00002

Table 3: Distributed Lying Cryptographers Example.

E More on the Multiple Lying Cryptographers Protocol

We revisit the Lying Cryptographers protocol from Section 6. The division into components is executed following the principles in Section 5. The hybrid approach divides the protocol into 8 components, one for each possible liar. Then each component is analyzed statistically. This allows the statistical analysis to sidestep the increased variance given by the presence of the liar, producing a more accurate result.

We can also use this example to show that LeakWatch does not compute bias correctly, and consequently produces an incorrect result. We have used QUAIL [28] to compute the exact Shannon leakage, and then compared the result of LeakWatch and of the hybrid approach. The results of the experiment are summarized in Table 3. Due to the bias correction problem explained in Section 4.1, LeakWatch computes an incorrect result: LeakWatch's bias correction only depends on the size of the joint matrix (as shown in Corollary 1), but the presence of zeroes in the matrix reduces the bias (as shown in Proposition 1). Finally, our hybrid approach manages to produce a result very close to the precise one computed by QUAIL, even if it uses statistical estimation.

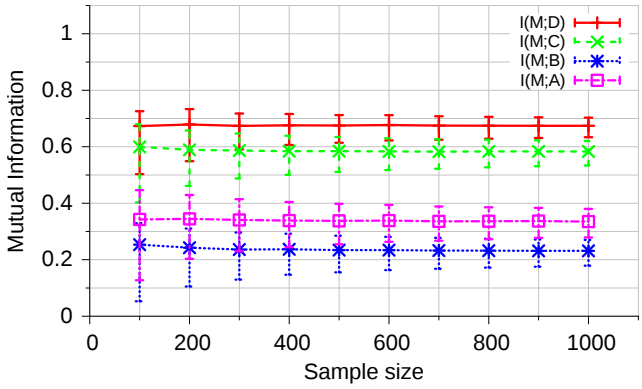
As regards previous tools on quantitative information flow, some comparison can be found in Biondi et al. [50].

F Example: Training of a Decision Tree

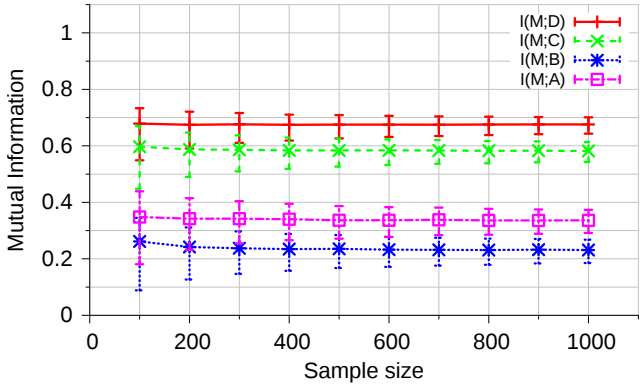
In this section we present an example outside the security scenario.

Decision trees are commonly used to build tree structures that classify categorical data. Consider having to take a decision that depends on multiple factors, for instance having to decide in which *malware category* a given malware is included. The malware category depends on whether the malware has certain *attributes*, for instance whether it sends the infected computer's data to an outside source, whether it installs a malicious driver, whether it displays advertisement, and so on. Since checking attributes for a given malware is expensive, we want to be able to categorize the malware by checking the minimal number of attributes. Let the attribute be from the set $\{A, B, C, D\}$, and the malware categories be from the set $M = \{$*Trojan, Spyware, Virus, Worm, Rootkit*$\}$. Each attribute is Boolean, so the malware can express it or not.

We will consider the ID3 algorithm [51] to construct a decision tree for malware classification. A decision tree is created, or "trained", by analyzing data from a database. In our case, each entry of the database associates a malware category with some of the attributes. For instance, malware of category *Worm* may have attributes A and C, while malware of category *Virus* may have A, C, and D. In this example we will consider that the ID3 algorithm has access to multiple databases, for instance because they are



(a) Fixed prior.



(b) Adaptive prior.

Fig. 9: Mutual information estimation for the decision tree training example (average of 1000 experiments).

provided by different sources. The information about malware on the different databases may overlap. We consider three databases with a size ratio of 1:3:6, which is known to the analyst.

The ID3 algorithm relies on being able to compute the mutual information between the target M and each of the attributes $\{A, B, C, D\}$. In particular, the root of the tree is chosen as the attribute with the highest mutual information with M. Since the databases are supposed to be very large and precise mutual information computation is expensive, this information is not freely available to the algorithm. Instead, we will show how to train a decision tree by statistically querying the databases and estimating the mutual information $I(M; X)$ between the malware categories M and each attribute $X \in \{A, B, C, D\}$ (with the associated confidence intervals).

F.1 Hybrid Statistical ID3 Implementation

To estimate the mutual information between the malware categories M and the attributes, we model each of the 3 databases as a joint probability distribution between M and each of the attributes $\{A, B, C, D\}$.

Ideally, assume that each malware datum in each database is associated with some unique ID i. The analyst randomly chooses an ID i and queries the database to obtain the malware category and attributes of the malware associated with i. The analyst repeats this many times to create the approximate joint distributions. Finally, they estimate the mutual informations $I(M; A)$, $I(M; B)$, $I(M; C)$, and $I(M; D)$ and the respective confidence intervals as explained in Section 3.

The root of the decision tree is the attribute $X \in \{A, B, C, D\}$ maximizing $I(M; X)$. However, due to the random sampling from databases, we need to gather enough data to claim with confidence that X actually maximizes mutual information. If two mutual information estimates have overlapping confidence intervals, we cannot guarantee that one of them has a value higher than the other. Hence, we will continue querying the databases until the confidence interval of the mutual information of the attribute with highest estimated mutual information does not overlap with any other confidence interval. Note that the ID3 algorithm normally requires precise analysis of the database: using statistical estimation and confidence intervals is our original approach. In this example we will compute only the root of the tree, due to space constraints.

F.2 Experimental Results

The results of the experiments are depicted in Fig. 9. We show the estimated mutual information values between M and each of the components and the relative confidence interval for increasing sample sizes. While the component with the highest mutual information is D, a certain sample size is required before its confidence interval is not overlapping with the confidence interval of the component with the second highest mutual information, i.e. C. In Fig. 9a we use a fixed prior on the different components, and the confidence intervals become non-overlapping when the sample size is greater than 800. On the other hand, in Fig. 9b we use the adaptive optimization of the priors presented in Proposition 3, and the confidence interval become non-overlapping when the sample size is greater than only 600 samples, proving the effectiveness of the adaptive optimization.