QASA 2012
(Pisa, Italy)

From Qualitative to Quantitative Information Erasure

Adedayo O. Adetoye    Michael H. Goldsmith

Cyber Security Centre,
Department of Computer Science,
University of Oxford,
United Kingdom
The Need for Information Erasure in the Context of Information Release

We want to process sensitive information but not necessarily propagate (some parts of) the information.

- **Statistical Databases**
  - May release sufficient information to be useful for statistical purposes, but must erase sufficient information not to violate privacy

- **E-commerce**
  - There are regulations on what information can be displayed by a merchant on receipts and screens, and what must be masked (*erased*)

- **E-voting**
  - We want to release result of election but not individual votes

...
Outline

1. Intuition about Information Erasure Modelling
2. Quantifying Information Erasure
3. Weaknesses and Advantages of Quantification
4. What Qualitative model does not capture
5. Conclusion
PCI stipulates that payment processing systems may display, on receipts and screens, at most first six and last four digits of CC Primary Account Number – other digits must be masked (erased).

Desired erasure policy is \( \forall c, c' \in \text{CC}. (c, c') \in R \iff c[1:6] = c'[1:6] \land c[13:16] = c'[13:16] \).
We consider deterministic systems

- Modelled as functions: \( f : X \rightarrow Y \)
- System’s input domain (contains secrets): \( X \)
- System’s output domain (the public observables): \( Y \)

Attacker can observe \( Y \) but not necessarily \( X \)

Attacker knows the system model \( f \)

How much of \( X \) is erased in \( Y \)?
Information Erasure: An Extreme

System modelled by $f : X \rightarrow Y$ erases all information (is noninterferring) if
\[ \forall x_1, x_2, f(x_1) = f(x_2) \]

\[ X \xrightarrow{f} Y \]

- The observation at $Y$ is independent of the choice of input $x \in X$
Intuition about Information Erasure: The Extremes

Figure: Extreme cases of Information Erasure
The level of erasure can be characterised by kernels of $f$, $g$, ... (or ERs)

(a) $\forall x, x' \in X, x \ all \ x'$ (total erasure of information in $X$)
(b) $\forall x, x' \in X, x \ id \ x' \iff x = x'$ (Input can be precisely determined from $g$ and $Y$)

Various other intermediate levels of information exist.
Example: Knowledge of Colour

∀ b, b' ∈ Balls. \( b \text{ col } b' \iff b.\text{colour} = b'.\text{colour} \)
∀ b, b' ∈ Balls. \( b \text{ all } b' \)

Figure: Information about colour
The ER model of information

Figure: Information about colour

all, col ∈ ER(Balls) are ERs over the set of Balls.

- A ER (R ∈ ER(X)) represents information by its ability to distinguish, or not, a pair of elements (x, x' ∈ X):
  - (x, x') ∈ R means indistinguishability of pair: lack of knowledge
  - (x, x') ∉ R means distinguishability of pair: knowledge
Suppose $R, R' \in ER(X)$

1. **Release Policy**: $R \rightarrow R'$
   - Given initial knowledge $R$, agent may not learn more than $R \cup R'$
   - Release because $R \subseteq R \cup R'$

2. **Erasure Policy**: $R \leftarrow R'$
   - Given some reference information $R'$, then $R \cap R'$ is the maximum allowed to be propagated
   - Or, if a system conforms to $R \leftarrow R'$ then it ensures that no more than $R \cap R'$ may be learnt from its output.
   - Erasure because $R \cap R' \subseteq R'$
A system modelled by $f$ satisfies the erasure policy $R \leftarrow R'$, written $f \models R \leftarrow R'$, if $\kappa_f \subseteq R \cap R'$.

Similarly, a system modelled by $f$ satisfies the release policy $R \rightarrow R'$, written $f \models R \rightarrow R'$, if $\kappa_f \subseteq R \cup R'$. 
Quantifying Erasure

Suppose \( R \in ER(X) \) and \( \mu \) is a probability measure over \( X \).

- The information content of \( X \) subject to its partitioning by \( R \) is
  \[
  \mathcal{H}(\mu|R) \triangleq - \sum_{X' \in [X]_R} \mu(X') \log \mu(X')
  \]
- A generalisation of the standard Shannon’s entropy definition
  \[
  \mathcal{H}(\mu) = \mathcal{H}(\mu|id_X) = - \sum_{x \in X} \mu(x) \log \mu(x)
  \]

Erasure and Release Quantification

\[
\begin{align*}
\mathcal{H}(\mu|R \leftarrow R') & \triangleq \mathcal{H}(\mu|R') - \mathcal{H}(\mu|R \cap R') \\
\mathcal{H}(\mu|R \rightarrow R') & \triangleq \mathcal{H}(\mu|R' \cup R) - \mathcal{H}(\mu|R)
\end{align*}
\]
Consider four systems modelled by the following functions:

1. \( f_1(x) = x \)
2. \( f_2(x) = |x| \)
3. \( f_3(x) = x \mod 2 \)
4. \( f_4(x) = 0 \)

**Figure:** Information Erasure and Release Policies
Duality of Erasure and Release

**Theorem**

For any chain of equivalence relations $R_1, R_2, R_3 \in ER(X)$ such that $R_1 \subseteq R_2 \subseteq R_3$ we have that

$$H(\mu|R_1 \to R_2) + H(\mu|R_2 \leftarrow R_3) = H(\mu|R_1 \to R_3) = H(\mu|R_1 \leftarrow R_3).$$

**Corollary**

For any set $X$, equivalence relation $R$ over $X$, and probability measure $\mu$ over $X$, we have that $H(\mu|\text{all}_X \to R) + H(\mu|R \leftarrow \text{id}_X) = H(\mu)$. 
Agrees with existing definitions

**Theorem (\(\mathcal{H}(\mu|all_X \rightarrow \kappa_f)\) equals mutual information)**

Let \(\kappa_f\) be the kernel of the function \(f : X \rightarrow Y\), then
\[
\mathcal{H}(\mu|all_X \rightarrow \kappa_f) = \mathcal{I}(X; Y).
\]
Furthermore,
\[
\mathcal{H}(\mu|\kappa_f \leftarrow id_X) = \mathcal{I}(X) - \mathcal{I}(X; Y).
\]

**Lemma (Erasure and Release between two comparable levels are identical)**

Let \(R, R' \in ER(X)\) such that \(R \subseteq R'\), and let \(\mu\) be a probability measure over \(X\). Then,
\[
\mathcal{H}(\mu|R \rightarrow R') = \mathcal{H}(\mu|R \leftarrow R').
\]
Caveat!

- Analysis requires $\mu$, what of if we don’t know it?
- Even with $\mu$, what does the measure mean?
Suppose $\mu(n) = \frac{1}{4}$ for all $n$.

\begin{align*}
\text{all} & : \\
0 & \quad 1 \\
2 & \quad 3 \\
\mathcal{H}(\mu|\text{all}) & = 2
\end{align*}

\begin{align*}
\text{Par} & : \\
0 & \quad 1 \\
2 & \quad 3 \\
\mathcal{H}(\mu|\text{Par}) & = 1
\end{align*}

\begin{align*}
\text{R} & : \\
0 & \quad 1 \\
2 & \quad 3 \\
\mathcal{H}(\mu|R) & = 1
\end{align*}

\begin{align*}
\text{id} & : \\
0 & \quad 1 \\
2 & \quad 3 \\
\mathcal{H}(\mu|\text{id}) & = 0
\end{align*}

**Figure:** Probability Permutation Problem of Quantitative Policies
Occlusion due to $\mu$ effectively restricting the function domain

Suppose $\mu(-2) = \mu(2) = \mu(-1) = \mu(1) = \frac{1}{4}$ and $\mu(-4) = \mu(4) = \mu(-3) = \mu(3) = 0$

$k_g$ where $g(x) = |x|

\begin{array}{cc}
-1 & 1 \\
-2 & 2 \\
\end{array}

\begin{array}{cc}
-3 & 3 \\
-4 & 4 \\
\end{array}

H(\mu|k_g) = 1

$k_f$ where $f(x) = x \mod 2$

\begin{array}{cc}
-1 & 1 \\
-2 & 2 \\
\end{array}

\begin{array}{cc}
-3 & 3 \\
-4 & 4 \\
\end{array}

H(\mu|k_f) = 1

Figure: Information Erasure and Release Policies
Information erasure is important in practice

We can model what information is erased in systems

Care should be taken with the interpretation of quantitative measures: what impact does \( \text{prob} \) (or our assumption about it) have on risk to information vis-a-vis the quantitative measure?

We may be able to constrain, via policies on \( \mu \), the probabilistic behaviour of systems and their environments as a statement of required system security to guarantee desired assurance

Many more interesting open issues: \textbf{Hybrid Qualitative + Quantitative Policies}, Reasoning about erasure of components of structured inputs, \textit{nondeterminism}, system composition and structuring ...
Thank You!

Questions?