Extending the Privacy Funnel for Nuclear Warhead Verification Protocols

Ruaridh Macdonald (rmacd@mit.edu), R. Scott Kemp
Department of Nuclear Science and Engineering, MIT
PI: R. Scott Kemp, rsk@mit.edu
Consortium for Verification Technology (CVT)

How can you compare different warhead verification protocols?

Many approaches to identifying nuclear warheads are under development. The variety of these efforts makes it difficult to directly compare protocols as their performance are assessed using different means and metrics.

We are developing a framework to systematically assess and quantify the accuracy and information protection of warhead verification protocols. This will enable like-for-like comparison of protocols.

The framework quantifies the completeness (type I error), soundness (type II error), and data protection of a protocol. These three values will assist protocol development as well as support inspector and nuclear warhead owner decisions in treaty negotiations.

What is mutual information?

Uncertainty about a variable can be described by calculating it's entropy: $H(X)$

$$H(X) = \sum_{X} -p(X = x)\log\big(p(X = x)\big)$$

Where $p(X = x)$ is the probability of the variable being equal to x.

$H(X)$ is reduced to zero after X is measured, as there is no longer any uncertainty as to its value. $H(X)$ is maximized when nothing about X is known and $p(X)$ is a uniform distribution.

Information about X can also be learnt by measuring a second, correlated variable. The reduction in the entropy of X after measuring a correlated variable Y is their mutual information: $I(X;Y)$[2]

$$H(X|Y) = H(X) - I(X;Y)$$

Acknowledgements

This work was partially funded by the Carnegie Corporation of New York

The authors would like to thank Kenneth Jarman (PNNL) for his significant contributions to the work.

References

1. Ali Makhdoumi, Salman Salamatian, Nadia Fawaz, and Muriel Medard. *From the information bottleneck to the privacy funnel*. In Information Theory Workshop (ITW), 2014 IEEE, pages 501–505. IEEE, 2014.
2. MacKay, David JC. *Information theory, inference and learning algorithms*. Cambridge university press, 2003.
3. R Scott Kemp, Areg Danagoulian, Ruaridh R Macdonald, and Jayson R Vavrek. *Physical cryptographic verification of nuclear warheads*. Proceedings of the National Academy of Sciences, 113(31):8618–8623, 2016.
4. Glaser, Alexander, Boaz Barak, and Robert J. Goldston. *A zero-knowledge protocol for nuclear warhead verification.* Nature 510, no. 7506 (2014): 497-502.

The Two-Party Privacy Funnel framework

The framework expands on data privacy methods [1] to model how questions about a test object are logically connected to the physical properties of the object and measurements which can be taken.

There are three types of information:

- **Latent / unmeasurable**
- **Private & measurable**
- **Public & measurable**

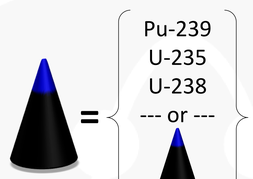
The completeness of the protocol is estimated by calculating the mutual information between the released data (Y) and the obfuscated data (X): **I(X;Y)**. The soundness is estimated by **I(C;Y)**.

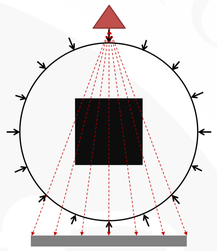
The maximum information that the inspector can learn about the object is **I(C;Y)**. This is agnostic to the inference method used.

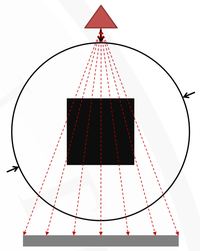
Note in some cases, this creates a soundness-privacy trade-off.

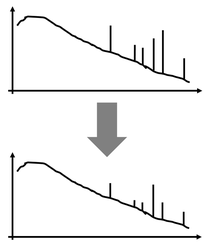
Both parties accept or reject the protocol based on their separate preferences. These preferences can be expressed as a function of I(X;Y) and I(C;Y), creating an optimization problem.

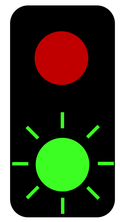
Calculating these mutual informations requires knowledge of the authentic warhead. That is possible for the owner but the parameters must be guessed by the inspector.

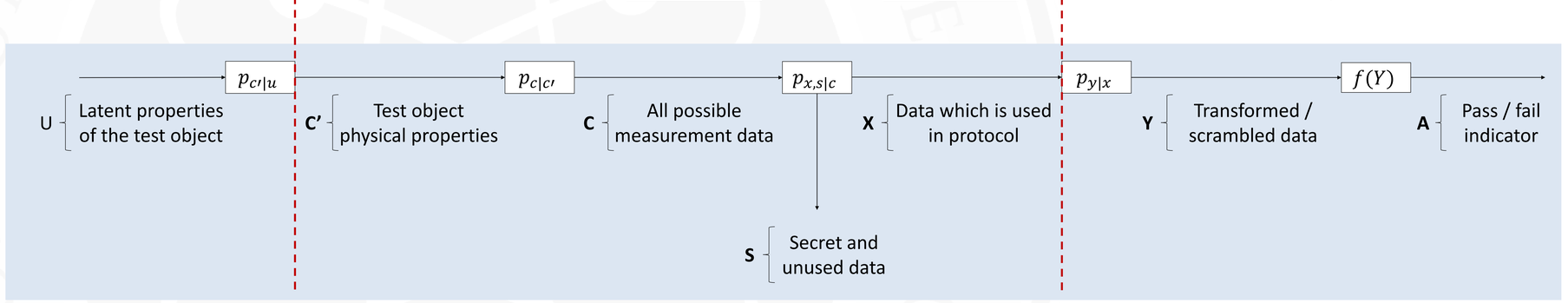




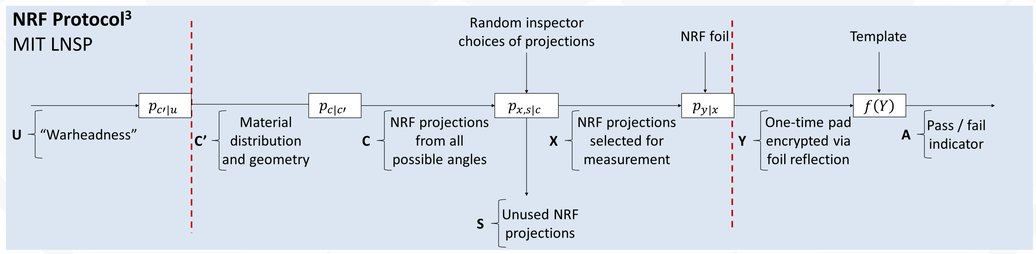


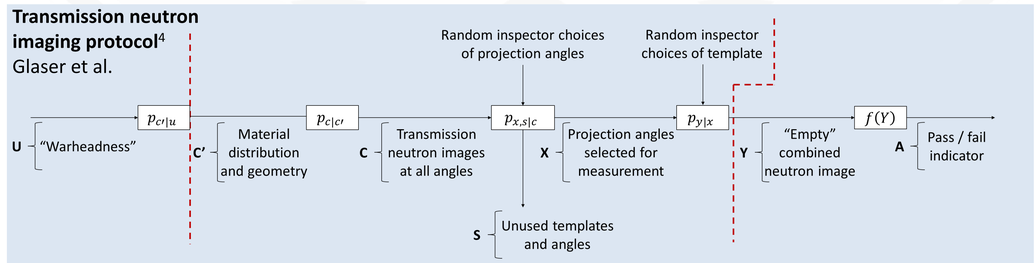






Examples





Legend:

- Information flow
- $p_{c\prime|u}$ — Axiom linking latent and physical properties
- $p_{c|c\prime}$ — Function linking physical properties to measurement results
- $p_{x,s|c}$ — Function which partitions potential measurements into used and unused sets.
- $p_{y|x}$ — Function describing the process by which the released data is obfuscated to protect warhead secrets
- $f(Y)$ — Function describing the comparison against a template or attributes

National Nuclear Security Administration