Designing Protocols for Nuclear Warhead Verification

Sébastien Philippe,* Boaz Barak,† and Alexander Glaser.*

*Nuclear Futures Laboratory, Princeton University, Princeton, NJ

†Microsoft Research, Cambridge, MA

ABSTRACT. Future arms-control and disarmament treaties could place numerical limits on all categories of nuclear weapons in the arsenals of weapon states, including tactical weapons, non-deployed weapons, and weapons awaiting dismantlement. Verification of such agreements is likely to require new types of inspection equipment — but also new verification protocols. This paper offers a set of definitions and building blocks to design verification protocols relevant to nuclear weapon authentication. It discusses how to construct and use physical interactive protocols with zero-knowledge property for inspections. The discussion illustrated by examples include topics such as perfect and statistical zero-knowledge, properties of the prover and the verifier, using trusted and non-trusted apparatus and detectors, physical commitment schemes and composition of zero-knowledge protocols.

Background

Future nuclear arms control and disarmament treaties may place limitations on all weapons and eventually require mechanisms to verify their dismantlement and disposition. Compliance with today's numerical limits is verified using indirect warhead counting techniques based on the easy accountability and identifiability of large strategic delivery vehicles. These techniques, that have been proven useful for strategic systems, will no longer be sufficient when addressing *all* nuclear weapons. New trusted and secure protocols for information exchanges and inspection activities must be ready in time to facilitate future negotiations and support the verification of baseline declarations.

This paper offers a set of definitions and building blocks to design verification protocols relevant to nuclear weapon authentication. It builds upon the cryptography literature with a focus on zero-knowledge proofs and the framework they provide to design inspections involving actual warheads.

This type of proof, where no knowledge is shared beyond the validity of a claim, was first introduced by Goldwasser, Micali, and Rackoff in the 1980s.[1] Since 2012, we have been applying these constructs, developed for digital application, to the non trivial

problem of nuclear weapon authentication by providing a physical zero-knowledge proof for warhead authentication using a template matching protocol.[2] Following this work, Fish, Freund, and Naor published the first attempt at a formal treatment of physical zero-knowledge proofs of physical properties using modern cryptography concepts.[3]

In this paper, we generalize the concept of zero-knowledge proofs to physical applications in a simple and useful way to facilitate their design and application in nuclear weapon inspections. The paper suggests definitions and discusses topics such as perfect, statistical and physical zero-knowledge, properties of the prover and verifier, trusted and non-trusted apparatus and detectors, physical commitment schemes and composition of zero-knowledge protocols.

Inspections with Trusted Third-Party

The difficulty of confirming the authenticity of a nuclear warhead can be summarized in a simple question. Can we convince someone of the assertion: "this object is a nuclear weapon", without giving away any knowledge beyond the fact that this assertion is true?

Traditional approaches to nuclear warhead verification have relied on engineered information barriers. An information barrier is "a system of procedural and technical measures designed to allow one or more unclassified measurements to be made on a classified object." In a more narrow sense, "an information barrier analyzes data that contains sensitive information and produces results that are then communicated as an unclassified output."[4] The fundamental challenge of the information barrier concept is the required *certification* and *authentication* of the equipment. Certification ensures the host that the device cannot reveal classified information and is safe to operate in the intended environment; authentication seeks to ensure the inspector that the device works as designed and displays genuine measurements.[5] Both hardware and software of the equipment have to be certified and authenticated.

In an ideal proof system, the information barrier plays the role of (or would be supplied by) a *trusted third-party* (TTP), which can enable secure and trusted interactions between two parties (host and inspector), similar to an escrow scheme. In the context of nuclear warhead verification, certification and authentication are the key procedures to establish the barrier as a trusted "third party." Detailed guidelines and procedures for the development and deployment of trusted information barriers have been proposed.[6] Ultimately, however, the critical questions for the overall viability of the concept remain: *Who has last custody of the equipment before it is used in an inspection?* And perhaps more importantly: *Who provides the critical equipment for the barrier?*

While acknowledging the advantages and disadvantages of both host-supplied and inspector-supplied equipment, project participants usually conclude that the informa-

tion barrier would be *de-facto* host-supplied. This makes certification relatively straightforward, but authentication extremely difficult.[7] From the outset, the inspector would therefore be at a disadvantage not only because the information barrier itself is provided by the host, but also because important authentication steps would have to be carried out in a host-controlled environment, perhaps using additional host-supplied tools and equipment.

We believe certification of inspector-supplied equipment may have more potential than is often assumed, i.e., it may be easier to resolve certification challenges than authentication challenges when both would have to be carried out in a host-controlled environment. This is discussed in an example further below. Simultaneously certifying and authenticating a trusted information barrier may be an elusive goal, however. For this reason, we place the main emphasis of our research on alternative approaches that do not require the use of engineered information barrier in the first place. To accomplish this task, we develop inspection protocols that are interactive zero-knowledge proofs.

Definitions for Physical Zero-Knowledge Proofs

A zero-knowledge proof is a proof that is both convincing to a verifier and at the same time, does not yield any knowledge but its validity. It is usually the result of interactions between a prover and a verifier where the verifier randomly challenges the prover whose responses convince eventually the verifier of the validity of his claim. No distinction can be made between actions the verifier can take after his interaction with the zero-knowledge prover, and actions he could have taken beforehand by simply believing the validity of the prover's claim.[8] Zero-knowledge is therefore a property of the prover only. It represents the prover's ability to resist attempts of a curious or malicious verifier to gain additional knowledge during the proof. All physical zero-knowledge proofs must be sound and complete. These are fundamental properties of all proof systems.

In a physical zero-knowledge proof of physical properties, the prover P wants to prove to the verifier V that an object O has a property X that P wants to keep secret (for example, X could be the presence of a sphere of plutonium and its radius would be the secret). To do so, P and V agree to participate in a protocol that will lead to the observation Y of X noted $Y|X$. Since the protocol is probabilistic (and measurements can be noisy), Y is a distribution rather than a single fixed value.

Definition 1. Perfect zero-knowledge. A physical proof of a physical property is *perfect* zero-knowledge if and only if for all X and X' that satisfy the property, the observations $Y|X$ and $Y|X'$ have the same probability law (i.e., are identically distributed), $Y|X = Y|X'$.

Because all physical objects that are manufactured from the same blueprints are likely to be different; two objects that are claimed to have identical properties may lead to different observations. However, these observations may be statistically indistinguishable from each other.

Definition 2. Statistical zero-knowledge. A physical proof of a physical property is *statistical* zero-knowledge if and only if for all X and X', the probability laws of the observations $Y|X$ and $Y|X'$ are statistically indistinguishable, $Y|X \approx Y|X'$.

Properties of a zero-knowledge proof. A zero-knowledge protocol guarantees that if the verifier behaves properly, then the prover won't be able to prove, except with small probability, a false statement. If the verifier doesn't follow the protocol, there is no guarantee on soundness - the prover may or may not be able to cheat and prove a false statement. On the other hand, if the prover behaves properly (and the statement is true), then the verifier will not learn any additional information. If the prover doesn't follow the protocol (and/or tries to prove a false statement), there is no guarantee on zero knowledge - the verifier may or may not be able to learn information.

The space of all zero-knowledge proofs for warhead authentication. Traditional approaches to nuclear weapon authentication have all relied on engineered information barriers to prevent sensitive information to be leaked during an inspection. We recognize that these have intended, maybe without realizing it, to be zero-knowledge proofs. It is certainly true that a warhead inspection system relying on a *trusted* information barrier that outputs a single bit observation $Y = \{Green, Red\}$ can be interpreted as a zero-knowledge proof. As we mentioned earlier, however, it has never been proven that engineered information barriers could be trusted by both the host and inspector parties.

Our approach, where we never measure sensitive information in the first place, proves that there exist relevant and interesting alternative members of the space of all zero-knowledge proofs for warhead authentication. It is possible to introduce classification of the different subspaces using, for example, the size and form of their observation Y. This is illustrated in Figure 2.

As we will see in the examples below, there is often a trade-off between the complexity of the equipment used to produce the observation and the size of the observation itself. A trusted information barrier may output a single bit observation (simple output) but can be a very complex measurement system (in terms of hardware and software). On the contrary, an inspection system using non-electronic detectors (simple measurement system) can output a more complicated observation and, if it is zero-knowledge, not leak sensitive information. By thinking about warhead inspections in zero-knowledge language we explore this trade-off.

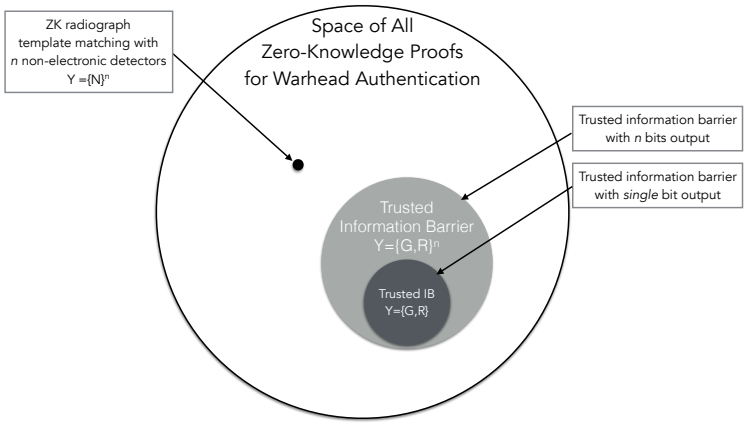

Figure 1: Representation of the space of all zero-knowledge proofs for warhead authentication. The proofs based on trusted information barriers are sub-spaces of the main space.

Designing Physical ZK Proofs and Examples

Treating nuclear weapon inspections as proofs provides a rigorous framework to compare the benefits and limitations of proposals to address this challenge. For example, all valid approaches should provide demonstration of their soundness, completeness and zero-knowledge properties.

To facilitate the construction of physical zero-knowledge proofs, Fish, Freund, and Naor proposed to separate the proofs in a logical layer and physical layer. All physical operations in the protocol belong to the physical layer. A rigorous mathematical treatment, including the demonstration of the zero-knowledge property, can then be done on an "hybrid-world" protocol where physical operations are replaced by their computational representations. This physics model based approach can be more or less complex depending on the requirements of the proof: for example, one can model a detector output by randomly sampling a known spectrum or can provide a full three dimensional physical model of a detector and, compute a spectrum based on its properties. In the first case the assumption is that there exists a detector that can provide a known spectrum, in the second that there exists a detector that has adequate dimensions and material composition.

If the "hybrid-world" protocol realizes an ideal zero-knowledge proof and is universally composable,[9] then any real-world physical protocol will conserve its zero-knowledge

properties as long as the most basic physical assumptions of the hybrid protocol are fulfilled.

Here we provide two examples for the design of protocols for warhead inspections and describe their logical layers. In both cases, we highlight necessary requirements so the proofs can be valid.

Example 1. *Attribute inspection with an information barrier provided by the inspector party.*

Both the host and the inspector have agreed on a set of attributes that defined a nuclear weapon. The host presents treaty accountable items in classified form for inspection. We start by outlining the logical layer:

One-Time Inspector Supplied Information Barrier: Logical Layer[10,11]

1. The inspector provides an information barrier that outputs a deterministic signature with a single bit only, $Y = \{0, 1\}$
2. The host takes custody of the barrier in presence of the inspector. If necessary, he runs a series of initial calibration tests.
3. The host proceeds with the inspection of treaty accountable items in the presence of the inspectors, gets the output and commits to them.
4. The host gains ownership of the device and runs any test he wants including destructive ones.
5. When convinced that the inspectors have provided an honest barrier, the host releases the output to the inspector.

Inspection barriers based inspections can only be considered to be valid proofs if the output observation Y can be trusted. This requires to provide physical detectors and barriers trusted by both parties. If the information barrier is provided by the host party, the inspector must be able to run program checks and calibrations measurement to trust the output. However, as we discussed before, this might not be enough if the host has placed hidden switches or Trojans in the software or hardware that the inspector is unaware of.[7,12]

Furthermore, since the information barrier will process secret information, the inspector will be unable to access it after the inspection is done. One way to address this asymmetry is for the inspector to provide the information barrier and limit the equipment provided by the host to physical analog objects (such as a detector crystals) that can be verified post inspection. Furthermore, if the host gains custody of the information barrier during the inspection and ownership of the information barrier after the

inspection (*One-time* information barrier), he can run theoretically an infinite number of checks and verifications of the software and hardware, including repeating the inspection and performing destructive measurement. If the information barrier is operated in a signal blocking glove-box, such as an RF enclosure and the observation generated by the information barrier is limited to one bit (Pass or Fail), the only way a malicious inspector could transmit secret information outside the box would be by designing the information barrier so that it can generate deterministic sequences of Passes and Fails. In this case, the device could in principle output an answer to another question (unknown to the host).

By making a commitment to the inspection results, the host can in theory verify that the inspectors have provided an honest box and eventually share the results with them.

Example 2. *ZK Proof for Radiographs Equality.*

A host wants to demonstrate to an inspector that two declared items A and B are identical by comparing their radiographs (using x-rays or neutrons). The host and the inspector agree that radiographs are a unique representation of an object, e.g., two significantly different objects cannot have the same radiograph.

ZK Proof for Radiographs Equality: Logical Layer

1. The host provides two radiographic films already exposed with the inverse image of object A and place then in two individual sealed envelopes.
2. The inspector *randomly* assigns object A and object B to one of the envelopes. The objects are placed between the envelope (at the image plane) and a radiation source.
3. The objects are exposed to the radiations. Both parties monitor independently the source fluence. This operation is essentially equivalent to adding a positive image on top of a negative image.
4. The host and inspector open both envelopes. Here a commitment strategy can be added if the host wishes to make sure that the results are not leaking knowledge due to his mistake (e.g., objects misalignment, wrong detector placement or source anomaly).
5. The verifier accepts or rejects the proof depending on the outcome: both images are flat gray background (items match their own negative) or there is a residual image appearing (items do not match their own negative).

If the inspector accepts the proof and the negatives are identical then item A and item B are identical. However the inspector doesn't know if the negatives are identical and can only compute probabilities. Since he or she takes a single random decision with two outcomes in the protocol, there is a 50 percent probability that the objects are indeed identical. On the contrary, if a residual image had appeared after opening the envelopes

(as it happens when a neagative to not match a positive), the inspector would have rejected the proof automatically.

The two parties can repeat the protocol n times to amplify the probabilities. Then the soundness (probability that the inspector will not accept two different objects as identical) and the completeness (probability that the host will convince the inspector that two identical objects are indeed identical) of the proof are both $1 - (1/2)^n$.

It is obvious that amplification by a large number of repetitions can be a costful and inadequate strategy, however it becomes unnecessary when a pool of objects needs to be proven identical (e.g., all weapons of a same type). For example, if we want to prove that 25 objects (labeled from B to Z) are identical to an object A. The host can prepare 26 negative images in advance. The inspector then randomly decides which negative will be used with which item. After all the inspections end, if all radiographs came out with a flat grey background, there is virtually no chances that the prover used different negatives for all items. If the host tries to conceal only one fake item out of 25, he has 96 percent chance to get caught and reveal design informations about the inspected objects.

Finally, the proof is zero-knowledge because the inspector does not learn anything beyond the result of the proof. To conserve the zero-knowledge property in the real-world physical protocol, the proof requires a non-electronic radiographic film or medium that can store information in multiple steps. In our application, we divided the radiograph in a grid and placed analog neutron detectors at every pixel and them designed an experiment to confirm the validity of the proof (see figure 2).[13]

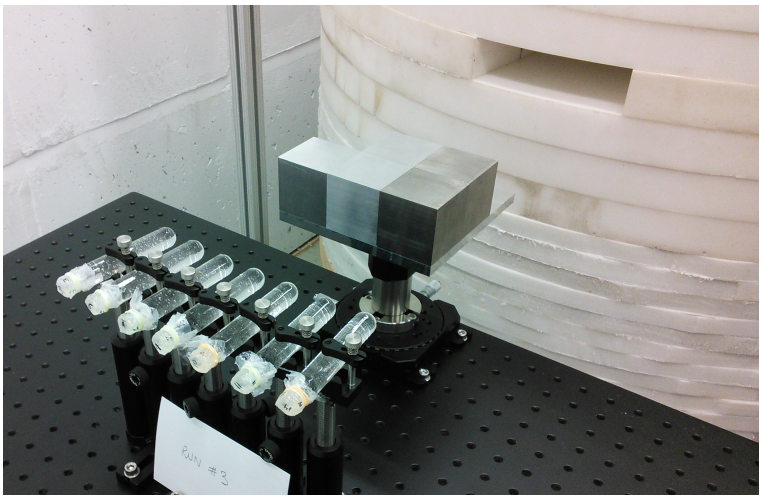

Figure 2: Photograph of the apparatus used in the first experimental demonstration of a zero-knowledge differential neutron radiographic protocol. (Image credit: S.Philippe).

Conclusion

Future arms-control and disarmament treaties could place numerical limits on all categories of nuclear weapons in the arsenals of weapon states, including tactical weapons, non-deployed weapons, and weapons awaiting dismantlement. Verification of such agreements is likely to require new types of inspection equipment — but also new verification protocols. Given that highly sensitive information has to be protected in the process, reference to and use of cryptographic concepts can offer valuable guidance in designing such protocols. In this paper, we have proposed a first set of definitions and building blocks, illustrated with specific examples, to start this discussion. Shared common definitions, concepts, and understandings of candidate approaches can hopefully facilitate peer review and collaborations between universities, non-governmental organizations and national laboratories.

The spectrum of possible proofs includes protocols that rely on an engineered information barrier that both the host and the inspector can simultaneously trust. Inspections following this principle are effectively equivalent to "trusted third-party" schemes in cryptography. At the other end of the spectrum are interactive zero-knowledge proofs that do not require trusted equipment because sensitive data is never acquired during the inspection (and thus need not be protected). In principle, both strategies could be successfully implemented, but there are tradeoffs in each case: Trusted third-party concepts involving information barriers can produce simple pass/fail signals at the expense of a potentially highly complex certification/authentication process. Interactive zero-knowledge proofs can offer inspection equipment, where simultaneous trust is much easier to establish (because sensitive information is not at stake), but they produce complex signals. A practical inspection system could in principle borrow and combine concepts from both trusted-third party schemes and interactive zero-knowledge proofs to produce robust measurement results while guaranteeing information security for the host party.

Endnotes

[1]S.Goldwasser, S. Micali and C. Rackoff, "The Knowledge Complexity of Interactive Proof Systems," *SIAM Journal on Computing*, Vol.18, pages 186–208, 1989.

[2]A. Glaser, B. Barak, and R. J. Goldston, "A Zero-knowledge Protocol for Nuclear Warhead Verification," *Nature,* 510, 26 June 2014, pp. 497–502.

[3]B. Fisch, D. Freund and M. Naor, "Physical Zero-Knowledge Proofs of Physical Properties," CRYPTO 2014, volume 8617, pages 313–336, Springer, Aug. 17–21, 2014.

[4]*Trust in Verification Technology: A Case Study Using the UK-Norway Information Barrier,* Non-Paper, NPT Review Conference, New York, May 2015.

[5]David Spears (ed.), *Technology R&D for Arms Control,* U.S. Department of Energy, Office of Nonproliferation Research and Engineering, Washington, DC, 2001, p. 7, www.fissilematerials.org/library/doe01b.pdf.

[6]UKNI, 2015, *op. cit.,* see in particular Figure 3.

[7]S. Philippe, A. Glaser, M. Walker, B. Barak and R. J. Goldston, "Resolving the Information Barrier Dilemma: Next Steps Towards Trusted Zero-Knowledge Nuclear Warhead Verification," INMM Information Analysis Technologies, Techniques and Methods for Safeguards, Nonproliferation and Arms Control Verification Conference, 12–14 May 2014, Portland, Oregon.

[8]O. Goldreich, S. Micali and A. Wigderson, "Proofs that Yield Nothing but Their Validity or All Languages in NP Have Zero-KNowledge Proof Systems," *Journal of the ACM*, Vol. 38, No. 1, pages 691–729, 1991.

[9]R. Canetti, "Universally Composable Security: A New Paradigm for Cryptographic Protocols," 42nd FOCS, pages 136-145, IEEE, Oct. 2001. Here by universally composable, we mean that the proof properties, including zero-knowledge, remain the same no matter how it is composed with other protocols including its execution in different physical environments.

[10]M. Kütt, A. Glaser, and S. Philippe, "Leveraging the Wisdom of the Crowd: Hardware and Software Challenges for Nuclear Disarmament Verification," *56th Annual INMM Meeting,* July 12–16, 2015, Indian Wells, California.

[11]B. Fisch, D. Freund, and M. Naor. Secure Physical Computation using Disposable Circuits. In Theory of Cryptography Conference 2015. https://eprint.iacr.org/2015/226.

[12]Becker G. et al., "Stealthy dopant-level hardware Trojans: extended version," *Journal of Cryptographic Engineering*, April 2014, Volume 4, Issue 1, pp 19-31.

[13]S. Philippe, R. J. Goldston, G. Ascione, A. Carpe, F. d'Errico, C. Gentile and A. Glaser, "Experimental Demonstration of a Physical Zero-Knowledge Protocol for Nuclear Warhead Verification," Proceedings of the 56th Annual INMM Meeting, Indian Wells, CA, July 2015.