A New Approach to Nuclear Warhead Verification Using a Zero-Knowledge Protocol

Alex Glaser,[*] Boaz Barak,[†] Rob Goldston[‡]

[*]*Princeton University*

[†]*Microsoft Research New England*

[‡]*Princeton Plasma Physics Laboratory and Princeton University*

ABSTRACT. Warhead verification systems proposed to date fundamentally rely on the use of information barriers to prevent the release of sensitive information. Measurements with information barriers significantly increase the complexity of inspection systems, make their certification and authentication difficult, and may reduce the overall confidence in the verifiability of future arms-control agreements. This article presents a concept for a new approach to nuclear warhead verification that minimizes the role of information barriers from the outset and envisions instead an inspection system that avoids the measurement of sensitive information, using a so-called zero-knowledge protocol. This is a protocol in which the data learned by one party (i.e., the inspector) allow him/her to verify that a statement is true (e.g., the inspected warhead is identical to an authenticated template), but does not reveal any additional information, e.g., does not leak any information that would help infer the design of the inspected warhead. There is a wide literature on zero knowledge proofs in the digital domain using cryptographic tools, and we draw on these ideas to achieve this in the physical domain. The proposed inspection system relies on active interrogation of a test object with 14-MeV neutrons, including both tomographic transmission measurements that are sensitive to warhead configuration, and scattering/fission measurements that are sensitive to material properties. The viability of the method is examined with MCNP Monte Carlo neutron transport calculations modeling the experimental setup.

Background

Existing nuclear arms-control agreements between the United States and Russia place limits on the number of *deployed* strategic nuclear weapons. Verification of these agreements can take advantage of the fact that deployed weapons are associated with unique and easily accountable delivery platforms, i.e., missile silos, submarines, and bombers. The next round of nuclear arms-control negotiations, however, may begin also to include tactical weapons and non-deployed weapons.[1] Both would require fundamentally new verification approaches, including authentication of nuclear warheads in storage and authentication of warheads entering the dismantlement queue. Dedicated inspection systems using radiation measurement techniques are likely to play a critical role in verifying such agreements, and different approaches have been proposed since the 1990s.[2] Independent of particular inspection approaches or measurement techniques there are general requirements for any viable inspection system:[3]

Certification: Before agreeing to use an inspection system, the host party will assure itself that the system does not divulge information that would be considered proliferation-sensitive or be otherwise classified. This process is called certification. Typical inspection systems considered to date fundamentally rely on the use of information barriers, which shield sensitive information from the monitoring party even though such information is in fact detected and processed by the system. The result of an automated analysis of the data can be displayed, for example, with green and red lights to indicate a passed or failed test.

Authentication: At the beginning of every inspection, even if based on a system that has been jointly developed, the monitoring party needs to assure itself that the particular instrument works as designed and that the data collected and displayed during the inspection are genuine measurements. This process is called authentication; it becomes necessary because the inspection system most likely remains with the host party prior to use, which may provide opportunities for tampering.

Completeness and Soundness: Even if a system can be certified and authenticated, both parties have to be confident that the inspection system meets the expectations and requirements of both parties. In particular, if a valid item is presented by the host, then the monitoring party will accept this item with high probability. Similarly, if an invalid item is presented, then the monitoring party will reject the item with high probability in spite of deception efforts that the host might undertake to defeat the system. In general, this will require extensive vulnerability assessments, which both parties can undertake independently in the design and development stages of the instrument.

In practice, two fundamental types of inspection systems have been distinguished, following either an attribute or a template-matching approach,[4] and both have advantages and disadvantages. The main challenge of the attribute approach is to establish and agree on meaningful properties of nuclear warheads. The method is also at a greater risk of spoofing; in fact, once threshold values for the attributes have been defined, the host knows exactly, which diversion scenarios will remain undetected. In spite of these shortcomings, the attribute approach has so far been the preferred method, in particular because it can avoid the use or presence of classified items prior to or during inspections.[5] For deeper cuts in the nuclear arsenals, however, more stringent verification requirements are necessary; in this case, authenticating warheads entering the dismantlement queue with the template-matching approach has important advantages because it offers the capability to distinguish warhead types and assures that no fissile material has been diverted. The challenge is to achieve this without revealing any information considered sensitive. This article summarizes first results of the Global Zero Warhead Verification Project, which has recently been launched in partnership with Princeton University's Nuclearfutures Laboratory,[6] to develop a new approach to warhead verification using template-matching method.

Experimental Setup and Conceptual Approach

Instead of working with mockups of fully assembled systems, this project envisions a simple experimental setup to demonstrate the basic concept. In our simulations, the test item is a highly simplified model of a plutonium pit consisting of a lead shell in a stainless steel enclosure. This test item is placed in the center of a carbon-steel drum. The drum itself is lined with medium-density fiberboard ("Celotex") with a central cylindrical cavity, in which the test item is held.[7] To interrogate the test item, we assume a 14-MeV neutron source, which will be available for experiments at the Princeton Plasma Physics Laboratory (PPPL) with a source strength of 1.5×10^8 neutrons per second emitted isotropically into $4\,\pi$. Neutrons are collimated by 40 cm of polyethylene, which is surrounded by a 10-cm steel sleeve to reduce gamma background, and illuminate the cavity for the test item in the container. The container is surrounded by a 270-degree cylindrical detector bank. In the computer model for the numerical analysis below, this array consists of 5,400 individual detector positions ($90{\times}60$ pixels, about one square-inch each).

Figure 1 shows the image of the test item that the surrounding detector bank would register, if we were not to use the zero-knowledge protocol described below. Two types of measurements can be distinguished: direct transmission measurements detecting 14-MeV neutrons, which produce a radiograph of the test item; and measurements at large angles, which detect scattered and fission neutrons and are particularly sensitive to material substitutions. Only the radiograph data are discussed in further detail below.

Mathematical

Direct observation of the data shown in Figure 1 would be unacceptable during an inspection carried out as part of a bilateral or multilateral verification arrangement because it reveals classified information about the test item. This dilemma has led to the concept and development of information barriers, which analyze the detected data

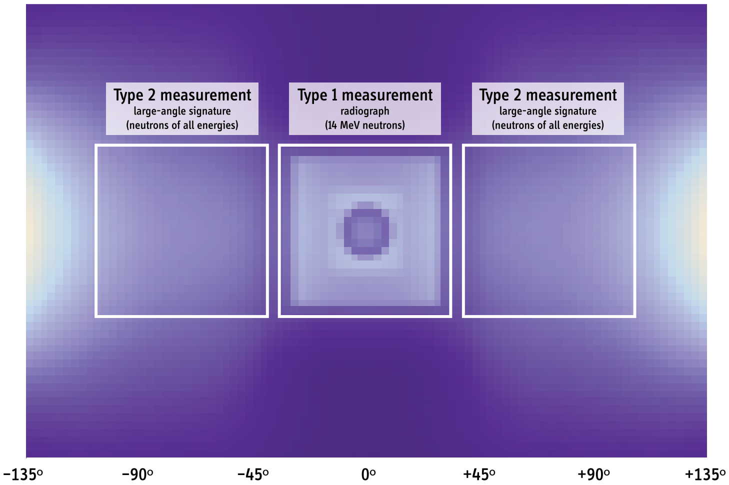


Figure 1: Transmission radiograph simulated with MCNP 5. The signal outside the direct field-of-view is dominated by fission and scattering from the sample. These data are never observed in our proposed approach.

and communicate the result with a simple yes/no answer. In general, an algorithm could compare the signatures of the test item and the template, and confirm their identity permitting certain tolerances and measurement uncertainties. Measurements with information barriers, however, significantly increase the complexity of inspection systems, making both their certification and authentication difficult. We avoid this problem by proposing a verification approach that permits full access to the data acquired in the measurement.

Typically, a proof of a mathematical statement not only assures us that a statement is true, but also explains *why* it is true. For example, the natural proof for the statement that a number C is not a prime number is to write down its factorization $C = P_1 \times P_2 \times \cdots \times P_n$, thus revealing not only the fact that the number is composite but also the actual factors. It turns out, however, that this proof—and in fact *any* proof—can be converted into a *zero knowledge proof*, i.e., a proof that does not reveal anything other than the validity of the original statement itself.[8] The crucial ingredients in obtaining such proofs are *randomness* and *interaction*. That is, rather than thinking of a proof as static text, we think of it as a protocol in which a prover and a verifier interact with one another. Moreover, the protocol involves tossing coins and, assuming that the protocol terminates successfully, statements are shown to be true with a specified confidence level, which can be set by the verifier.

In typical cryptographic applications, zero knowledge proofs are used to prove mathematical statements in the digital domain. They have found many applications including in privacy-preserving data mining, electronic voting, and online auctions.[9] Here, we explore the opposite approach and propose a hardware implementation of a zero-knowledge protocol for warhead verification. To illustrate the general idea, consider the following setting, which is closely related to our proposed approach for warhead verification:

> *Alice (the prover) has two small bags each containing the same number of marbles. She wants to prove to Bob (the verifier) that both bags contain the same number of marbles (detector counts), without revealing to him what this number is. To do so, Alice prepares in advance ten pairs of buckets. For each pair, she pours the same large but randomly chosen number of marbles in both buckets of the pair. Alice presents all pairs to Bob, who chooses nine out of the ten pairs and examines them, verifying that indeed for in each of these nine pairs both buckets contain the same number of marbles. Bob has now high confidence that the last pair of buckets also contains an identical number of marbles, though he does not know what this number is. Alice pours the marbles from the first bag into the first remaining bucket and the marbles from the second bag into the second bucket. Bob examines both buckets and accepts the proof if and only if they both contain the same number of marbles. In principle, this protocol can be repeated as many times as Bob wishes.*

More formally, one can show that participating in the protocol yields Bob arbitrarily little new information about the number of marbles in the bags. In a Bayesian setting, Bob's beliefs about the number of marbles in a bag—or the design of a nuclear weapon—will remain the same before and after interacting with Alice.

Results of Monte Carlo Neutron Transport Simulations

Even if zero-knowledge proofs can be shown to exist for any mathematical problem, it is by no means clear that such protocols can be effectively implemented in meaningful ways for real-world applications. The possibility of using zero-knowledge protocols for nuclear warhead verification would be particularly valuable given the sensitivities involved in this application. The following discussion focuses on fundamental measurement concepts that might offer the possibility of zero-knowledge rather than on particular detector technologies that might be used for such an inspection system.

As shown in Figure 1, active neutron interrogation is particularly useful for measurements on fissile materials and many complex measurement techniques including neutron coincidence and multiplicity measurements have been developed to probe fissile material configurations.[10] Following here the most basic approach, we consider signatures based on the total neutron counts only, which—if proven adequate—would significantly reduce the complexity of the inspection system. To be specific, assume that a neutron detector has registered a certain neutron count N during the time allotted to the measurement, for example, in decimal and binary notation:

$$N = 137\,531 \quad \rightsquigarrow \quad \underbrace{0\,0\,0\,1\,0\,0\,0\,0\,1}_{\text{drastic changes}}\overbrace{1\,0\,0}^{\text{bits}}\underbrace{1\,0\,0\,1\,1\,1\,0\,1\,1}_{\text{statistical noise}} \tag{1}$$

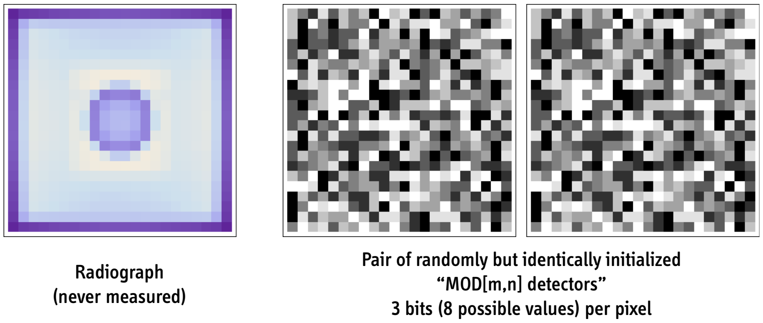
The statistical error of the measurement is proportional to $\sqrt{N}$ and does not contain any useful information for comparing the test item to the template. These least significant bits are not further considered. Sensitive design information or drastic differences between the test item and the template would be observable in the most significant bits of the number; these too are not further considered. The most valuable bits for our purposes are those that are above the statistical noise, above permissible differences that might be due to manufacturing tolerances, and above variations in detector sensitivity after calibration. For the sample analysis below, we retain three selected bits of the total detector count, i.e., 8 possible numerical values, but other choices are possible and could be equally effective.

Another important element of the proposed approach is to pre-initialize pairs of detectors with random noise, generating so-called one-time pads.[11] As an example, and as illustrated in Figure 2, the detector arrays that will be used for the radiograph measurement each consist of 22×22 pixels, which are pre-initialized *identically but randomly,* i.e., in the this example with 3 binary bits, or values between 0 and 7, represented by shades of gray in the figure.

During a typical measurement, the selected bits will be switched often; in fact, depending on the detector position and the choice of bits, possibly up to hundreds or thousands of times. As an example, assume that the initial detector state was 1 0 1 and that the measurement itself (without initialization) would have resulted in 1 0 0 as would be the case for the sample count rate above. The actual value registered by the detector is 0 0 1 as the overflow bit is not registered in the measurement:

$$1\,0\,1 \oplus 1\,0\,0 = [1]\,0\,0\,1 \quad \text{or in decimal notation:} \quad (5+4) \bmod 8 = 1 \tag{2}$$

Using this hypothetical detector design, complete measurements on the template and test items can be simulated using MCNP 5 neutron transport simulations. Since the



Radiograph
(never measured)

Pair of randomly but identically initialized
"MOD[m,n] detectors"
3 bits (8 possible values) per pixel

Figure 2: Prior to the inspection, a pair of detector arrays is initialized with random but identical values. In this example, three bits will be retained in the actual measurement, which corresponds to eight possible shades of gray. The radiograph of the test item is shown for reference purposes only and is never recorded during the inspection following the proposed protocol.

detectors are initialized with a one-time pad, and overflow bits are not recorded, the post-measurement detector states remain completely random (Figure 3, top). Nothing can be learned from analyzing the state of individual detector arrays and, by extension, nothing can be learned about the test item.

Despite their randomness, detector states resulting from measurements on the template and the test item can be compared and their similarity quantified. To this end, patterns are compared using a "shortest-distance" operation. For example, assume that one detector displays $1\,1\,1$ and its counterpart $0\,0\,0$; there are two ways (Δ_1 and Δ_2) to go from one to the other:

$$1\,1\,1 \ominus \underbrace{1\,1\,1}_{\Delta_1} = 0\,0\,0 \quad \text{and} \quad 1\,1\,1 \oplus \underbrace{0\,0\,1}_{\Delta_2} = [1]\,0\,0\,0 \tag{3}$$

In other words, the shortest distance between and $1\,1\,1$ and $0\,0\,0$ is $\Delta_2 = 0\,0\,1$, which corresponds to $(7+1)$ mod $8 = 0$. These comparisons can be performed for every pixel of both patterns. In the case of a match, i.e., in the case of a test item that is identical or quasi-identical to the template, a comparison of two post-measurement detector states will only reveal residual statistical noise (Figure 3, bottom left).[12] If significant differences between the template and the test item exist, however, the comparison will reveal those as also shown in the figure below. Figure 4 shows results for some additional diversion scenarios, all of which can be identified with the proposed approach.[13]

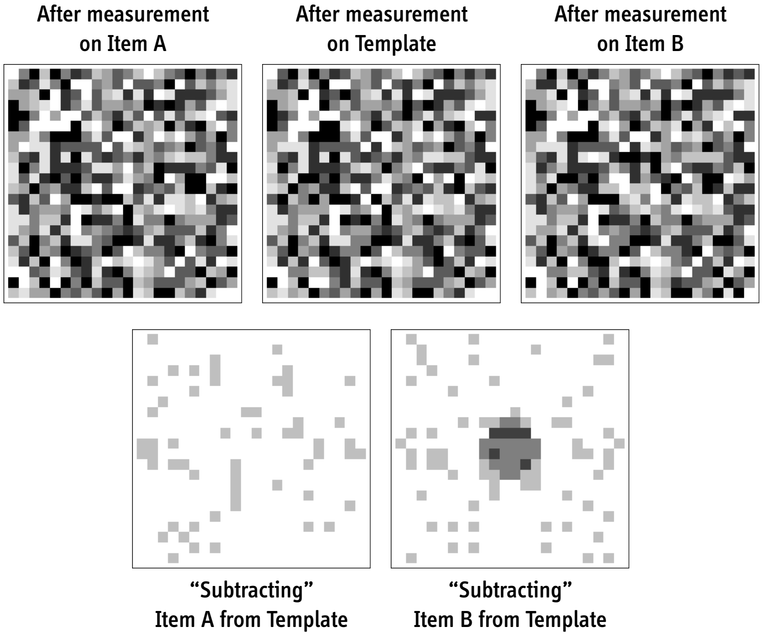

After measurement on Item A
After measurement on Template
After measurement on Item B
"Subtracting" Item A from Template
"Subtracting" Item B from Template

Figure 3: TOP: Status of detector arrays after measurements on a valid item (left), on the template (middle), and on an invalid item (right). As before exposure, all patterns are random, and inspectors can have full access to them. BOTTOM: After subtracting a pair of patterns, the data can be used to distinguish valid from invalid items. If a valid item is presented (left), the patterns are equivalent and only statistical noise is present; if an invalid item is presented, statistically significant differences appear. In this case, 800 grams of plutonium have been removed from the pit. 14 MeV neutrons, MCNP 5 simulations, 10 billion source neutrons.

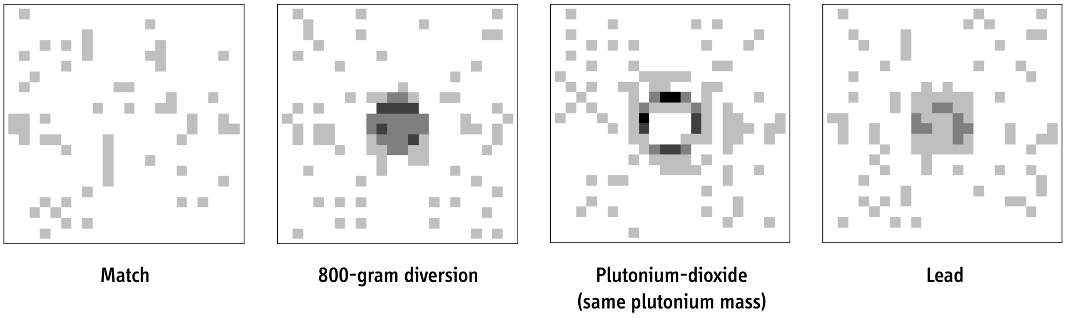


Figure 4: Different types of diversion scenarios can be distinguished using the proposed zero-knowledge protocol using modular arithmetic. MCNP 5 simulations.

Conclusion

Dedicated inspection systems are likely to play a critical role in verifying future arms control agreements, which may cover both tactical and non-deployed nuclear weapons and require verified warhead dismantlement. A major new verification challenge will be to authenticate nuclear warheads offered for inspection without divulging classified information. Using so-called information barriers is one possibility to accomplish this task, but such barriers result in complex inspection systems that are difficult to certify and authenticate. Here we have proposed a fundamentally new approach to nuclear warhead verification using the template-matching method combined with a tool from modern cryptography, i.e., a zero-knowledge protocol, where nothing can be learned about the inspected item from the data acquired in the measurement other than that it is quasi-identical to a reference item. Data sets are random but differences between them, should they exist, are revealed when comparing two sets. Preliminary results based on Monte Carlo neutron transport simulations suggest that many important diversion scenarios can be identified with a hypothetical inspection system based on this approach.

Overall, zero-knowledge protocols appear as an important new approach to nuclear warhead verification. If implemented effectively, they could reduce the relevance of information barriers and result in inspection systems that are easier to certify and authenticate. These concepts and technologies need to be developed now in order to be available for the next round of arms-control negotiations.

Acknowledgement. *This project is supported by a generous grant from Global Zero, funding from Princeton Plasma Physics Laboratory, and in-kind contributions from Microsoft Research New England. This work is also supported in part by DOE Contract DE-AC02-09CH11466.*

Endnotes

[1] For example, U.S. President Obama noted in March 2012: *"Going forward, we'll continue to seek discussions with Russia on a step we have never taken before—reducing not only our strategic nuclear warheads, but also tactical weapons and warheads in reserve."* Remarks by President Obama at Hankuk University, Seoul, Republic of Korea, 26 March 2012.

[2] For an older overview, see, for example, David Spears (ed.), *Technology R&D for Arms Control,* U.S. Department of Energy, Office of Nonproliferation Research and Engineering, Washington, DC, Spring 2001, www.ipfmlibrary.org/doe01b.pdf.

[3] The discussion below partially follows the terminology established by the Authentication Task Force Report, jointly developed by the U.S. Departments of Energy and Defense, Washington DC, June 2001.

[4] Under the attribute approach, the host and the inspecting party agree on a set of unclassified properties, or attributes, of a nuclear warhead or warhead components and seek to develop a system that can confirm these attributes in a yes/no manner, i.e., without revealing any quantitative information. For example, such a system could confirm that more than a certain minimum mass of a particular fissile material is present. In contrast, under the template approach, a measurement generates a complex and unique fingerprint of the inspected item. This fingerprint is then compared against the fingerprint of a reference item, or template, to confirm that both items are substantially identical. The template has been previously confirmed to be authentic.

[5] For example, the 1996–2002 Trilateral Initiative between Russia, the United States, and the IAEA only confirmed the presence of plutonium, a minimum mass, and constraints on the isotopics. It did not require or involve any items in classified form. The parties could either present actual plutonium pits or ingots produced from them.

[6] See www.globalzero.org and nuclearfutures.princeton.edu for more details.

[7] Overall, this configuration is similar to the AL-R8 container, which is used in several U.S. facilities to store plutonium pits. Luisa F. Hansen, *A Comparison of the Shielding Performances of the AT-400A, Model FL, and Model AL-R8 Containers,* UCRL-JC-120849 (Preprint), Lawrence Livermore National Laboratory, U.S. Department of Energy, April 1995.

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

[9] Bernard Chazelle, "The Security of Knowing Nothing," *Nature,* 446, 26 April 2007.

[10] *Technology R&D for Arms Control, op. cit.*

[11] The "one-time pad" encryption encrypts a secret message $x \in \{0, 1, \ldots, c\}^n$ with a key $k \in \{0, 1, \ldots, c\}^n$ by outputting $x + k \pmod{c}$, i.e., the i^{th} digit of the ciphertext is equal to $x_i + k_i$ if $x_i + k_i < c$ and is equal to $x_i + k_i - c$ otherwise. If the key k is chosen uniformly at random then, regardless of the value of x, the ciphertext is random and hence reveals no information about the message. However, if the same key is used for more than one message then information will leak, therefore the name "one-time pad;" this is also the reason why in practice people prefer (conjecturally) computationally-secure encryptions such as the Advanced Encryption Standard (AES) over the one-time pad or other information-theoretically secure encryption schemes.

[12] Identical objects are compared in MCNP simulations using different random seeds at the beginning of each simulation.

[13] More challenging are substitutions involving only isotopics of the fissile material, e.g., replacing weapon-grade with reactor-grade plutonium. In this case, additional measurements at large angles (as indicated in Figure 1), which detect scattered and fission neutrons originating from the test item and are particularly sensitive to material substitutions, may become necessary.