Information Analysis Technologies, Techniques and Methods for Safeguards, Nonproliferation and Arms Control Verification Workshop

Workshop Proceedings

May 12-14, 2014
Portland Marriott Downtown Waterfront Hotel
Portland, OR, USA

Co-Sponsored by:
Pacific Northwest Chapter of the Institute of Nuclear Materials Management
INMM Division of International Safeguards
INMM Division of Nonproliferation and Arms Control

INMM
INSTITUTE OF NUCLEAR MATERIALS MANAGEMENT

»Information Analysis Technologies

Contents

Resolving the Information Barrier Dilemma:
Next Steps Towards Trusted Zero-Knowledge Nuclear Warhead Verification

Sébastien Philippe,[1] Alexander Glaser,[1] Mark Walker,[1]
Boaz Barak,[2] and Robert J. Goldston[3]

[1] *Nuclear Futures Laboratory, Princeton University*
[2] *Microsoft Research*
[3] *Princeton Plasma Physics Laboratory and Princeton University*

ABSTRACT
In the cryptography literature, zero-knowledge protocols are interactive proof systems where a prover can demonstrate the validity of an assertion to a verifier, who will accept the proof with a high probability while not gaining any knowledge beyond the validity of the prover's claim. Going at counter-current from previous historical development, we translated the concept of zero-knowledge proofs first developed for digital applications into the physical world by proposing an approach to nuclear warhead authentication envisioning an inspection protocol that a priori avoids detection of sensitive information. Under such physical zero-knowledge protocol, the host (prover) can prove to an inspector (verifier) that a warhead is authentic without revealing anything about its materials or design. The current developments focus on the practical implementation of such system using non-electronic detectors. After discussing the advantages of both the zero-knowledge and non-electronic properties of our authentication system compared to alternative proposals using so-called information barriers, we present some computational test-results and their impact on the design of our experimental proof-of-concept.

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 strategic bombers. The next round of nuclear arms-control negotiations, however, may begin also to include tactical weapons and non-deployed weapons. 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 to accomplish this task.[1] These are usually divided in two categories commonly referred as the template or attribute approaches. The so-called template method is generally considered the most robust verification approach. It envisions the comparison of a complex fingerprint of an inspected item against the fingerprint of a reference item, or template, to confirm that both items are substantially identical. In contrast, the attribute method verifies that the inspected item correctly features a limited set of previously agreed attributes with the consequences of providing a rigid framework where cheating may become easier as well as inherently revealing warhead design information.

What makes the authentication of nuclear weapons difficult in the framework of bi- or multilateral inspections is that any radiation measurements made on these objects, as well as any physical means (hardware) to process these measurements would themselves be highly classified. In the case of the United States, it is important to recall, “nuclear weapons knowledge is born secret”.[2] This greatly limits

the ability of an honest host (prover) to convince the inspector party (verifier) that the object presented is in fact an authentic warhead, *completeness* of the proof, without compromising any classified information. On the other hand, it is equally difficult for the inspector to be certain that the verification procedure cannot be tricked into accepting false or tampered objects, *soundness* of the proof, without being able to see the measurement data and its associated processing equipment. Both completeness and soundness are in this case greatly affected by the classified nature of the object being inspected.

To address this problem, all proposed inspection systems have had, so far, to rely on engineered information barriers to protect classified information at the cost of introducing inherent limitation in the completeness and soundness of these protocols. Yet none of these have demonstrated the ability to achieve both *perfect information security* for the host and *perfect information integrity* for the inspector.

To free ourselves from this conundrum, we have proposed a fundamentally different approach based on the cryptographic concept of *zero-knowledge proofs.*[3] Using a zero-knowledge protocol eliminates the need of an information barrier in the first place. Going further, we address the important problem of untrusted hardware/electronics and software in authentication system by using non-electronic (pre-loadable) detectors only.

This paper discusses the advantages of both the zero-knowledge and non-electronic properties of our warhead authentication system in the framework of information security and integrity. A series of computational test-results are presented and their impacts on the design of our experimental proof-of-concept are briefly described.

GOING NON-ELECTRONIC

Information barriers are often depicted as complex technological systems using electronic hardware components and potentially sophisticated software. They require both parties to trust that they have no trapdoors hidden from the inspector, which could be used to cause a system to declare invalid object as authentic, nor side channels unknown to the host, which could leak classified information. This raises the issue of the potential presence of untrusted hardware and malicious software within information barriers. Any piece of electronic or line of code can potentially be hacked and tampered.

Ideally, the development of such barriers is carried out jointly between the host and inspector parties to build confidence in the overall security architecture of the system.[4] However, the host is still likely to supply all the building components as well as upload any software prior to measurement. There exists no enforcement mechanism to prevent the host from conducting secret parallel development with the goal of tampering equipment or exploiting any vulnerability he may identify, including after the development phase has been completed. All security checks during post-silicon validation and post-manufacturing testing would therefore be particularly difficult for the inspector party wishing to ensure that the hardware meets the design specifications and that it works as designed.

Hardware Trojans are one important example illustrating the inherent difficulty of hardware authentication. These Trojans are defined as malicious changes or additions to an integrated circuit that add or remove functionality or reduce reliability of the system.[5] The host can potentially insert hardware Trojans at many stages in the product cycle: in the design stage, the manufacturing stage, the assembly stage, and the shipping (supply-chain) stage.

In the design stage, joint development on all aspects of the integrated circuit design from defining high-level functions down to wiring at the transistor level can prevent the host from including malicious side channels. The inspector can then check if the manufactured circuits correspond to the agreed design blueprints. If the information shared during the development phase is limited to high-level functions, while the low-level circuit design remains unspecified,[6] then the inspector would need to reverse-engineer the hardware using recently developed techniques.[7] If all building blocks of the information barriers can be selected by the inspectors at the time of the inspection, (e.g. by cut-and-choose one circuit out of ten), and if the remaining parts can later be (destructively) analyzed, then the swapping of circuits during the assembly and shipping stage should not be an issue.

Trojan inclusion during the manufacturing stage probably remains the hardest host strategy to detect and deter. Material trojans, based for example on aging or environmental (i.e. temperature) effect on transistors, are extremely difficult to detect. The inspector may not be able to recreate all the variables including environmental ones necessary to trigger the hardware modification when testing the device independently.[8] Recent work has shown that hardware can be even tampered at the sub-transistor level by modifying the dopant masks.[9] Since only changes to the metal, polysilicon or active area can be reliably detected, these dopant Trojans remain immune to detection today.

Finally, once a measurement on a classified item is complete, the inspector would generally not be allowed to re-inspect the equipment. This constraint limits the ability to trust completely the "green light" given beyond the information barrier, even for a system that is assembled with components selected in a cut-and-choose manner because, using the example of the hardware trojan, the non-expected system behavior may only be triggered (temporarily or permanently) during the inspection. Despite these challenges, strong information barriers can be developed as the example of the UK-Norway Initiative has shown.[10] Nevertheless, it may prove difficult or perhaps impossible to convince "truly skeptical" hosts and inspectors of the viability of an inspection system (used in a treaty context on real warheads or warhead components) given the inherent limits to security and integrity of the electronic components used in such a system.

For these reasons, we decided to examine the use of a non-electronic detection mechanism based on physical phenomena and permitting post-measurement inspection. To perform template authentication, when electronic real-time read-out cannot be used, non-electronic detectors must have the primary ability to *store* data, for example, the total number of neutron counts, in order to be read and analyzed at a later stage. Two detector technologies meet these criteria: superheated emulsions ("bubble detectors") and neutron activation analysis detectors (e.g. cylinders or prisms of zirconium). These detector types have also important properties in line with the requirement of our zero-knowledge protocol. They have the capability to be preloaded with a desired neutron count prior to the inspection. This preload can persist for hours or days and its decay or aging rate, if present, is well characterized. Preload counts are indistinguishable from counts accumulated during irradiation of the test items. The detectors are energy selective so that the effect of low energy neutrons returning from room walls in an experiment can be minimized. They are insensitive to gammas, have relatively high efficiency, and permit total counts in the range of several thousands to tens of thousands. Other technologies may have similar properties but we will give these the highest priority for our experimental proof-of-concept.

In superheated emulsions, neutron recoil particle can trigger the formation of macroscopically observable bubbles from microscopic droplets that are dispersed in an inert matrix.[11] If the detector vials are stored in adequate conditions to prevent aging, bubbles can be kept in steady state for years.

Gel formulation with high droplet density can be achieved. Temperature control is an important factor for these detectors as both the energy threshold and response of the emulsion are temperature sensitive. Adequate control of environment variables during experiment is therefore important.

Activation analysis is used in fusion research.[12] The concept is based on using energy threshold reaction type such as (n,n') and (n,2n) reactions with significant cross sections. Product isotopes of these reactions eventually decay emitting gamma radiations. These are consequently counted in adequate detectors after irradiation. Potential candidates for our experiment include ^{90}Z(n,2n)^{89}Z with a 12 MeV threshold and ^{115}In(n,n')^{115m}In with a 1 MeV threshold. These types of detector prevent the host from visually inspecting any preload however they would require shielding. As opposed to bubble detectors, counting decay events may require hours or even days. However they have the ability to hold a much larger number of counts.

If no electronics is used in detectors, certain parts of the experiment apparatus will still require to be controlled electronically – the neutron generator and post processing equipment among others. These are believed to be much less sensitive since they do not hold, measure or store any classified information. Therefore the inspectors may access them for extensive study upon request. Finally it is important to stress that our non-electronic detectors are neither considered sensitive or classified after the measurements campaign. This is related to the zero-knowledge knowledge property of our authentication protocol and is an important difference with existing methods.

THE VIRTUES OF KNOWING NOTHING

Zero-knowledge protocols are *interactive* proof systems where a prover can demonstrate the validity of an assertion to a verifier, who will accept the proof with a *high probability* while not gaining any knowledge beyond the validity of the prover's claim. In other words, after the proof protocol ends, the verifier gains no new knowledge about the object of the proof while being convinced of its validity. The notion of zero-knowledge remains a property of the prover and is not affected by the behavior of the verifier even when he intends to cheat.[13]

This powerful concept, first developed in the *digital* domain, can have interesting non-trivial *physical* application. In the subject of our interest, we argue we can prove two nuclear warheads are identical or *similarly close* using a zero-knowledge template approach without revealing any information whatsoever on their design, composition or other sensitive data.

To illustrate the concept of physical zero-knowledge proof, we present a simple example related to our authentication protocol:

Alice (the host) has two small cups both containing X marbles where X is some number between 1 and 100. She wants to prove to Bob (the inspector) that both cups contain the same number of marbles, without revealing to him what this number X is. To do so, Alice prepares two buckets, which she claims each contain (100 – X) marbles. Bob now randomly chooses into which bucket which cup is poured. Once this is done, Bob verifies that both buckets contain 100 marbles.

This simple protocol reveals no information on X since, regardless of its value, Bob should always counts 100 marbles in both buckets. If the cups did not have the same number of marbles, then no matter how the buckets are prepared, with probability of 50% after the pouring, one of the bucket will not contain

100 marbles. If Alice and Bob repeats the game five times then if Alice is consistently cheating, she will be caught with $(1 - 2^{-5}) > 95\%$ probability.

Similarly we design a zero-knowledge protocol for warhead authentication. We want to compare a candidate item with a template warhead by recording their direct transmission pattern of 14-MeV neutrons, as well as the intensity of neutrons emitted to the side of the items at a typical angle of 90 degree. In analogy to the marble example, the measurements are recorded using *non-electronic* detectors that are *preloaded* with the negative image of the radiograph. The protocol is as follow:

1. The host presents to the inspector the two items to be compared as well as two sets of preloaded non-electronic detectors. Preloaded values are not revealed to the inspector.

2. The inspector chooses which set of detectors will be used with which item.

3. Once the radiography is done. The inspector verifies that the final images are uniform.

If the items actually differ, and the preloads are chosen to complement the two items, then with significant probability of at least 50% the image will not be uniform.

As opposed to the marble example, neutron measurement is inherently statistically noisy. To avoid conveying information through the noise, the detectors are preloaded with noisy values. Since both the preload and the measurement distribution will follow a Poisson distribution and since the sum of two Poisson distribution with mean and standard deviation $(N_p, \sqrt{N_p})$ and $(N_m, \sqrt{N_m})$ is also Poisson with $(N_p + N_m, \sqrt{N_p + N_m})$, our protocol achieves the following:

The neutron count obtained by any measurement on the template or on a valid submitted item is distributed according to the Poisson distribution with mean and variance equal to a previously agreed value N_{max}.

With N_{max} being chosen in advance by both sides, neither the measurement nor its noise reveals any new information. N_{max} could correspond for example to the maximum number of counts that is expected in the absence of a test item. If a submitted item varies from the template (or the submitted preloads are not identical) an image may be seen that could contain sensitive information. This will be an additional strong incentive for the host not to cheat. However, it also means that the system in its most simple form may be prone to human error. This might necessitate the addition of a fail-safe mechanism to prevent the host for reveling information when committing an unintentional mistake (i.e. wrong disposition of preloaded detectors). Finally, the steps following a measurement should be relatively straightforward. Since the information contained in the detectors is in principle unclassified, protocols can be devised that permit using both host-provided and inspector-provided measurement tools.

COMPUTATIONAL ANALYSIS

We now show our approach can be implemented in practice and that it can detect some notional diversions between the two objects. The current setup (see Figure 1) features an array of 367 super-heated droplet neutron detectors, a deuterium-tritium (D-T) neutron generator, placed in a polyethylene collimator. Items to be authenticated are placed between the neutron generator and the

detector array. Bubble detectors have the particularity to be sensitive above a certain energy threshold. Transmission detector are assumed to be sensitive to neutron energy above 10-MeV. Side detectors are currently getting developed and are likely to feature an energy threshold of 1-MeV in order to capture potential fission and other compound nucleus reaction neutrons.

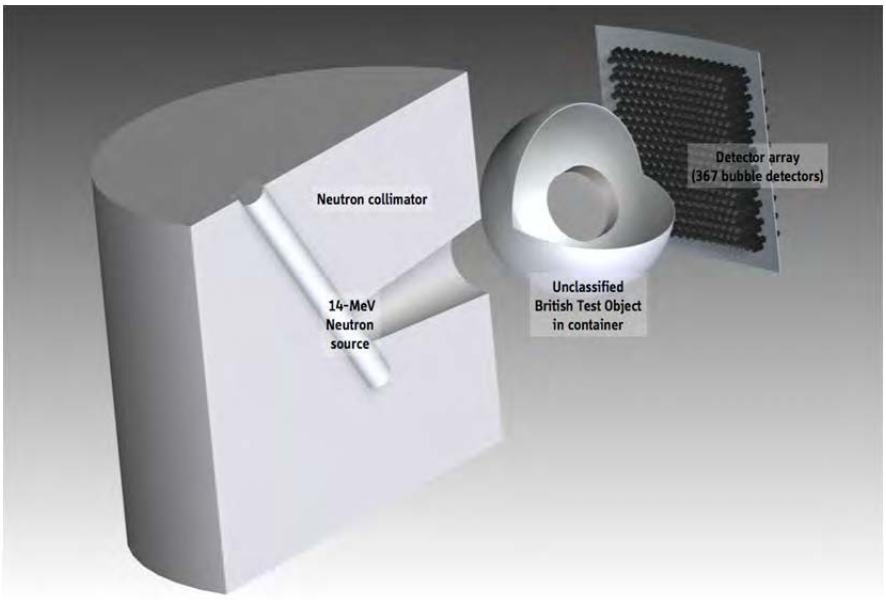

Figure 1. Experimental setup with neutron source in collimator, test item in container, and detector array. Large-angle detectors are not shown.

Figure 2 illustrates the transmission results on a valid and invalid item. The sensitivity of the measurement scenarios increases with N_{max} and the associated improvements of counting statistics.

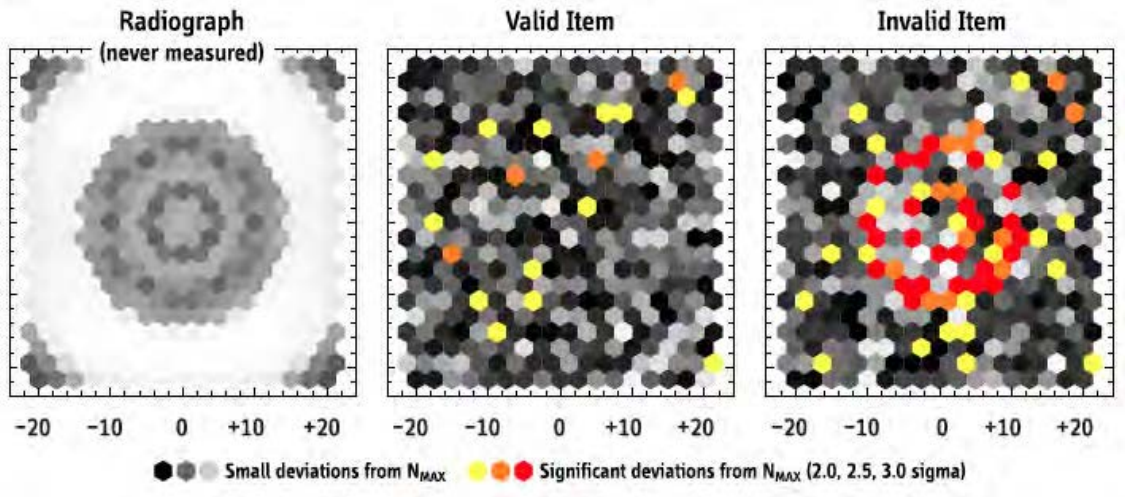

Figure 2. Results of MCNP5 simulations illustrating valid and invalid items.[14] The radiograph of the test

item shown on the left is never measured, i.e., corresponds to a measurement without preloading the detectors. The other panels show total detector counts after measurements on a valid and an invalid item. Shades of gray and colors indicate absolute differences from $N_{max} = 1000$. The invalid item produces a larger number of suspicious data points, which are in this case spatially correlated.

Additional simulations were performed to determine the range of possible detection for the diversion scenarios of material removal and substitution in both large and small (spatially localized) quantities. We found that the required N_{max} for large removal or substitution (for example tungsten versus lead) of materials can be as low as 1000. However local substitution of material, for example the replacement of 7% of a total mass of about 8 kg of tungsten in the British Test Object by lead may require a total count of 32,000 to achieve a detection probability of 95%.

These numbers are nevertheless conservative since they do not take into account the use of side detectors. The more realistic case of substitution of uranium-238 for uranium-235 in a nuclear weapon component results in a factor of about two reduction in the induced fission rate due to 14 MeV neutrons. Substitution of reactor-grade for weapon-grade plutonium has a small effect on the directly induced fission rate, but a large effect on the spontaneous fission rate, which could be detected passively by operating the side detectors in the absence of the neutron source.

CONCLUSION

Authenticating nuclear warheads without revealing classified information represents a qualitatively new challenge for international arms-control inspection. Here we have shown an example of a zero-knowledge protocol based on non-electronic differential measurements of transmitted and emitted neutrons that can detect small diversions of heavy metal from a representative test object. This technique will reveal no information about the composition or design of nuclear weapons when only true warheads are submitted for authentication. It therefore does not require an engineered information barrier. The use of non-electronic detectors prevents the most sensitive equipment of our authentication system to be vulnerable to untrusted hardware and software.

If the computational results are encouraging there is still a lot to achieve before a practical system can be readily implemented. Current research focuses on the design and construction of the first experimental proof-of-concept. Topics include shielding of the neutron source, development of the side detectors, effect of room returns on measurement. All have effect on the minimum achievable N_{max} for the various diversion scenarios.

Timely demonstration of the viability of such an approach could be critical for the next round of arms-control negotiations, which will likely require verification of individual warheads, rather than whole delivery systems. Other such zero-knowledge protocols may be possible. The zero-knowledge approach has the potential to remove a major technical obstacle to new nuclear arms control agreements that include both deployed and non-deployed, strategic and tactical weapons, at substantially lower levels of armament than current agreements.

Acknowledgement. This project is supported by generous grants from Global Zero and the U.S. Department of State, funding from the Princeton Plasma Physics Laboratory, supported by DOE Contract

DE-AC02-09CH11466, and in-kind contributions from Microsoft Research. All simulations were run on Princeton University's High Performance Cluster.

---

[1] David Spears (ed.), Technology R&D for Arms Control, U.S. Department of Energy, Office of Nonproliferation Research and Engineering, Washington, DC, Spring 2001.

[2] Peter Galison, Removing Knowledge, Critical Inquiry, vol. 31, Iss. 1, 2004, pp. 229-243.

[3] Alexander Glaser, Boaz Barak and Robert J. Goldston, *Towards a Practical Implementation of an Inspection System for Nuclear Warhead Verification With an Inherent Information Barrier*, 54th Annual INMM Meeting, 14-18 2013, Palm Desert, California.

[4] United Nations, The 2010 Review Conference of the Parties to the Treaty on the Non-Proliferation of Nuclear Weapons, *The United Kingdom-Norway Initiative: Research into the Verification of Nuclear Warhead Dismantlement*, NPT/CONF.2010/WP.41 (May 2012), available from undocs.org/NPT/CONF.2010/WP.41.

[5] Becker G. Intentional and Unintentional Side-Channels in Embedded Systems. PhD dissertation submitted to the Graduate School of University of Massachusetts Amherst. February 2014.

[6] There are many low-level circuit designs that will produce the same high-level functionality.

[7] Subramanyan, P *et al.* "Reverse Engineering Digital Circuits Using Structural and Functional Analyses," *IEEE Transactions on Emerging Topics in Computing*, early access article (2014). And: Randy Torrance and Dick James, The State-of-the-Art in IC Reverse Engineering, Cryptographic Hardware and Embedded Systems – CHES 2009, Lecture Notes in Computer Science Volume 5747, 2009, pp. 363–381.

[8] Sharad Malik, personal communication, May 2014.

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

[10] David M. Chambers et al., "UK-Norway Initiative: Research into Information Barriers to Allow Warhead Attribute Verification Without Release of Sensitive or Proliferative Information," INMM 51st Annual meeting, Baltimore, MD, USA, July 11-15, 2010.

[11] d'Errico, F. Radiation dosimetry and spectrometry with superheated emulsions. *Nuclear Instruments and Methods in Physics Research B* 184, 229-254 (2001).

[12] Bleuel, D. L. *et al.* Neutron activation analysis at the National Ignition Facility. *Reviews of Scientific Instruments* 83, 10D313 (2012).

[13] Oded Goldreich, Foundations of Cryptography, 1st ed. Vol. 1, Cambridge: Cambridge University Press, 2001.

[14] A general Monte Carlo N-particle (MCNP) transport code. Los Alamos National Laboratory, mcnp.lanl.gov.