Chapter 4

DETECTING CYBER ATTACKS ON NUCLEAR POWER PLANTS

Julian Rrushi and Roy Campbell

Abstract This paper proposes an unconventional anomaly detection approach that provides digital instrumentation and control (I&C) systems in a nuclear power plant (NPP) with the capability to probabilistically discern between legitimate protocol frames and attack frames. The stochastic activity network (SAN) formalism is used to model the fusion of protocol activity in each digital I&C system and the operation of physical components of an NPP. SAN models are employed to analyze links between protocol frames as streams of bytes, their semantics in terms of NPP operations, control data as stored in the memory of I&C systems, the operations of I&C systems on NPP components, and NPP processes. Reward rates and impulse rewards are defined in the SAN models based on the activity-marking reward structure to estimate NPP operation profiles. These profiles are then used to probabilistically estimate the legitimacy of the semantics and payloads of protocol frames received by I&C systems.

Keywords: Nuclear plants, intrusion detection, stochastic activity networks

1. Introduction

Digital instrumentation and control (I&C) systems are computer-based devices that monitor and control nuclear power plants (NPPs). Analog I&C systems have traditionally been used to perform monitoring and control functions in NPPs. However, Generation III+ and IV reactors are equipped with digital I&C systems; meanwhile, analog systems in older reactors are being replaced with digital systems. In general, NPP control networks communicate with SCADA systems to coordinate power production with transmission and distribution demands. The deployment of digital I&C systems and the connectivity between NPP control networks and external networks expose NPPs to cyber attacks whose consequences can include physical damage to reactors [3].

Please use the following format when citing this chapter:

Rrushi, J. and Campbell, R., 2008, in IFIP International Federation for Information Processing, Volume 290; *Critical Infrastructure Protection II*, eds. Papa, M., Shenoi, S., (Boston: Springer), pp. 41–54.

In addition, security exercises have underscored the threat of intrusions into NPP control networks.

This paper focuses on the problem of intrusion detection in NPP control networks. It addresses the following problem: Given a set of protocol data units (PDUs) received by a digital I&C system over an NPP control network, how could the digital I&C system assess if every PDU is legitimate and is not a component of attack traffic?

Digital I&C systems considered in this paper communicate via the Modbus protocol [8]. However, the intrusion detection approach is applicable to any industrial protocol. The approach falls in the category of protocol-based and anomaly-based intrusion detection strategies. Such an approach guards against attacks that use semantically regular PDUs to target an NPP as well as attacks that use irregular PDUs (e.g., memory corruption attacks).

The first set of attacks can take an NPP to abnormal conditions and initiate physical damage. NPPs are equipped with reactor protection systems that monitor operational variables and shut down systems if pre-defined thresholds are passed; this reduces the risk of physical damage. Nevertheless, the attacks impact availability because it takes several hours to restart an NPP. The second set of attacks, on the other hand, evade reactor protection systems and have the potential to cause physical damage, mainly because several NPP sensors are shared between I&C systems and reactor protection systems. Thus, attackers can initiate physical damage to an NPP as well as defeat reactor protection systems by passing them fake status data.

2. Operation-Aware Intrusion Detection

This paper proposes operation-aware intrusion detection as an anomaly-based defensive capability for NPPs. Profiles of legitimate behavior of digital I&C applications are constructed by analyzing payloads of PDUs sent over control and/or fieldbus networks and the semantics of each PDU field in the context of NPP operations. Thus, PDUs are analyzed in terms of the bindings between streams of bits and tasks such as withdrawing control rods, changing the reactor feed pump rate or closing steamline isolation valves. This approach provides more visibility than model-based detection with regard to the potential of PDUs being legitimate or malicious. As a matter of fact, PDUs may be perfectly formatted according to the protocol specifications while having the potential to cause harm to an NPP.

For example, well-formatted PDUs that close the main turbine control valves when reactor power goes above 25% take the reactor to an anomalous state in which the reactor vessel pressure rises far beyond the maximum allowable value. A turbine bypass valve is supposed to provide for excess steam flow, but the flow capacity of the bypass valve is normally equivalent to 25% of steam flow. This flow capacity is insufficient to return the reactor vessel pressure to normal soon enough to maintain safe NPP conditions. Furthermore, several application-level attacks can be crafted that comply with protocol specifications. For example, memory addresses and shellcode injected as malicious PDUs in Modbus memory

corruption attacks could be made to appear as valid coil and/or holding register values.

Modbus memory corruption attacks [1] highlight the need for operation-aware protocol-based intrusion detection. These attacks exploit faulty mappings between the addresses of data items defined by Modbus and the memory locations where the data items are actually stored. Register addresses and values used in such attacks are generally fully compliant with Modbus specifications. Nevertheless, it is unlikely that the attack PDUs will consistently have valid semantics in terms of NPP operation. These operational aspects of PDUs and the associated NPP states are leveraged by the intrusion detection approach described in this paper.

3. Modeling NPP Operations

The stochastic activity network (SAN) formalism [5, 11, 14] as implemented by the Möbius tool [2, 12] is used to model and analyze the operation of a boiling water reactor (a common type of NPP) along with its digital I&C systems that engage the Modbus protocol. The operation of Modbus applications in digital I&C systems, the control of digital I&C systems over NPP components and the interactions among NPP components as reflected by NPP operational variables are captured using atomic SAN models. Figure 1 presents an excerpt of an atomic model developed for a digital I&C system. Discrete inputs, coils, input registers and holding registers are modeled as SAN places, which we call "device places." The number of tokens in each device place represents the value of the Modbus data item that corresponds to the SAN place.

Depending on its configuration, a Modbus device may have as many as 65,536 data items; the corresponding atomic model would be extremely large. To address this issue, the information contained in PDU data fields is modeled as a set of SAN places, which we call "PDU places." The numbers of tokens in PDU places represent the values placed in the data fields of PDUs sent to digital I&C systems. PDU function codes are modeled as instantaneous activities, e.g., write coil, write multiple coils, write register and write multiple registers. Some of these instantaneous activities have several cases. The probability distributions of case selections for these instantaneous activities depend on the markings of the input places associated with the input gates of the instantaneous activity under consideration.

In general, case selections indicate the data items accessed from Modbus device memory. PDU places are fed with tokens by timed activities (not shown in Figure 1 due to space limitations) that model the behavior of a master device generating Modbus requests. The enabling predicates in the input gates in Figure 1 check whether the input places associated with the gates are not all zeroed, in which case the corresponding instantaneous activities are enabled. (A PDU that is not sent over the network is modeled using PDU places containing zero tokens.) Upon completion of an instantaneous activity, one of the associated activity cases is selected. In the case of instantaneous activities, when modeling a write function code of any kind, the output functions of out-

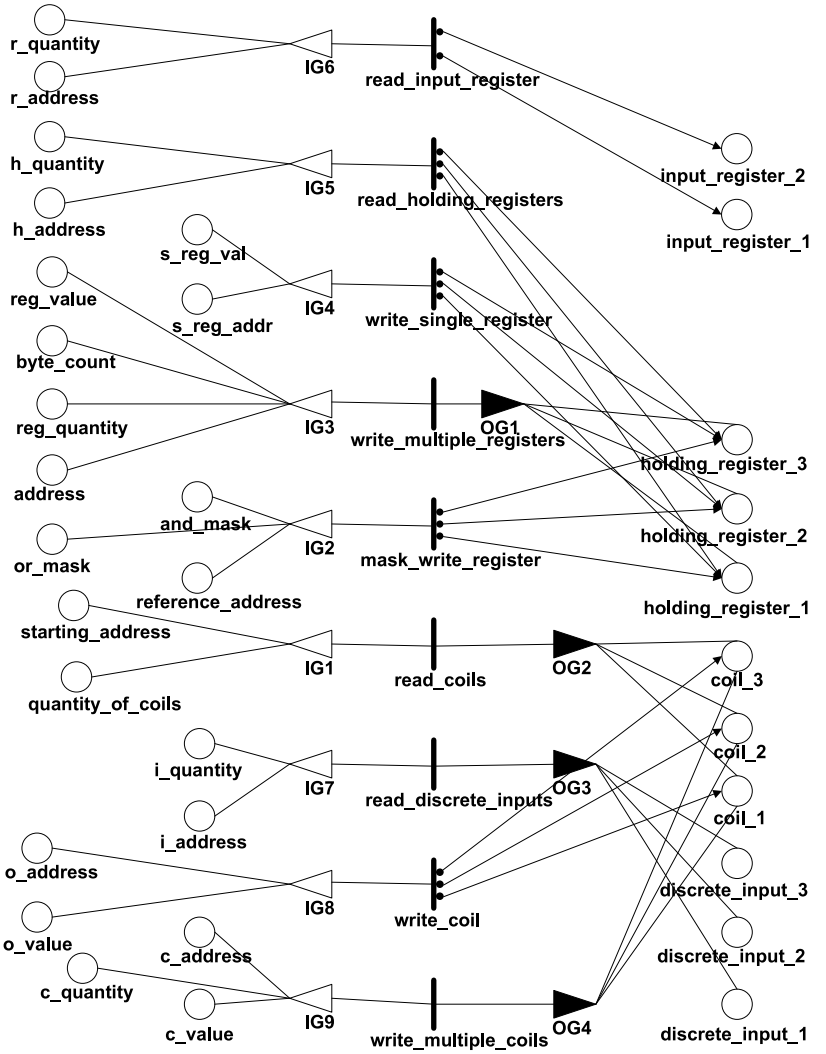


Figure 1. SAN model excerpt for a Modbus digital I&C system.

put gates associated with the selected activity case add a number of tokens to one or more device places.

The output functions zero the PDU places that enable the instantaneous activities modeling function codes associated with the output gates in question. In the case of instantaneous activities, when modeling a read function code, output functions act as identity functions on device places and add tokens to a set of places modeling variables for a master device (not shown in Figure 1 due to space limitations). The variables related to the operation of NPP physical components are also modeled as a set of places, which we call "NPP

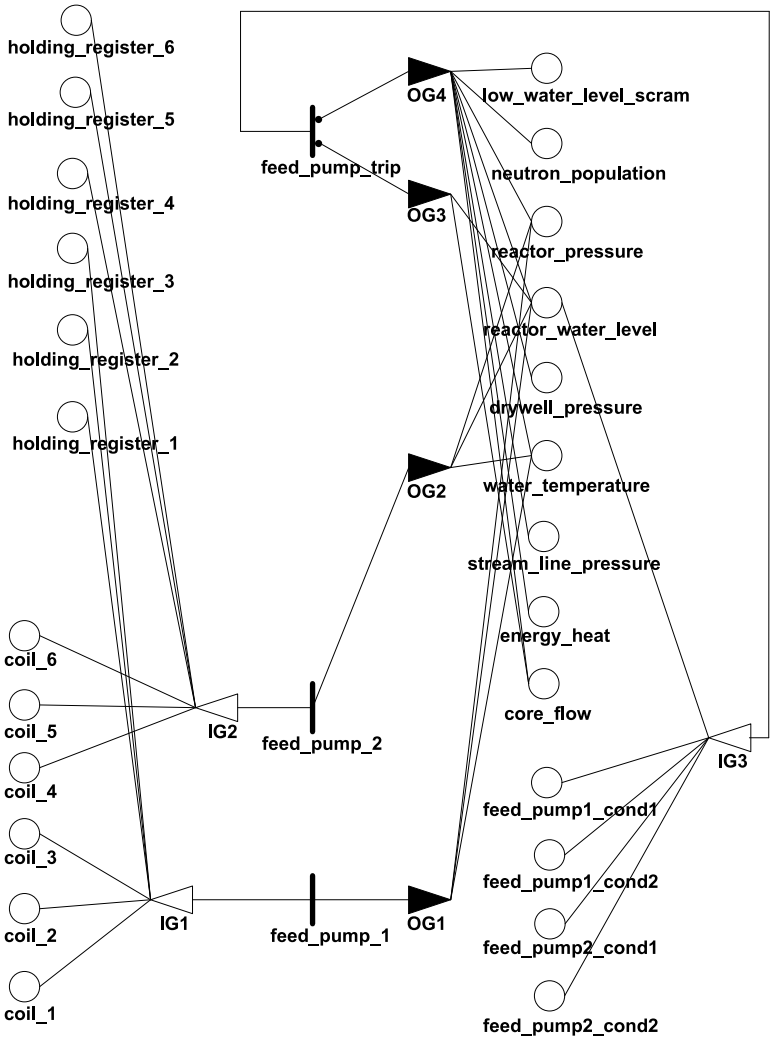


Figure 2. SAN model excerpt of a digital I&C system with two reactor feed pumps.

places." Examples include temperature and pressure of water and steam for a component such as the NPP moderator. The number of tokens in an NPP place represents the value of the NPP operational variable modeled by the place. For example, the fact that the reactor pressure is 900 psi is modeled as an NPP place (reactor pressure) that holds 900 tokens.

Figure 2 presents a SAN model excerpt of the operation of a digital I&C system that controls two reactor feed pumps. The mechanisms used to operate on the NPP are modeled as timed activities. The two reactor feed pumps are examples of these mechanisms, which, in terms of NPP operations, raise the pressure of the moderator to make it flow to the reactor vessel. The rates of

these timed activities are generally obtained from the technical specifications accompanying the physical equipment modeled by the timed activities and specific configurations of the Modbus devices that control the equipment.

For example, the rates of the timed activities feed_pump_1 and feed_pump_2 are in the range [0, 8.4]. The maximum capacity of the reactor feed pumps modeled by the timed activities is 8.4 million lb/hr. The rates of these timed activities are calculated based on the markings of the device places acting as input places associated with the input gates of the timed activities. The markings considered are those that hold at the moment the timed activities are enabled. The enabling predicates of the input gates associated with feed_pump_1 and feed_pump_2 check if the numbers of tokens in the input places are changed by comparing them with the numbers of tokens in the set of places holding old values (not shown in Figure 2 due to space limitations).

The output functions of output gates associated with cases of timed activities (e.g., OG1 and OG2 in Figure 2) add a number of tokens to one or more NPP places and mark as old the numbers of tokens in the input places of the input gates associated with the timed activities. Correctly modeling transients in a SAN is crucial to constructing profiles of normal NPP operations. Transients are abnormal events (e.g., loss of feed water for heating, inadvertent initiation of a coolant injection system, pipe breaks, and mechanical or electrical faults) that perturb key NPP operational variables. In most cases, NPP operation can be returned to normal conditions without initiating a "scram" – a rapid shutdown of the NPP.

NPPs are required to maintain redundant systems, which are activated in the event of failure to prevent transient conditions from escalating into accidents. Sometimes transients are caused by unrecoverable faults that may result in a reactor scram initiated manually by NPP operators or automatically by the reactor protection system. In the nuclear power environment, transients are "normal events" that are expected to occur during NPP operations. Our SAN model expresses transients as timed activities (e.g., the feed_pump_trip timed activity in Figure 2 models the loss of one or two reactor feed pumps). From the intrusion detection point of view, transients must be treated with care so that they do not become sources of false positives.

Each transient has specific effects on NPP operational variables. Perturbations to NPP variables caused by a transient are reflected as changes to the values of data items stored in the memory of Modbus devices. However, there is always uncertainty whether drastic changes to data items during NPP operation are caused by transients or attacks. We employ the activity time distribution functions of timed activities that model possible transients to distinguish between transient effects and attack effects. Note that the estimation of the failure rates of physical NPP components is outside the scope of our work since it requires the consideration of electrical, chemical and mechanical properties of NPP components.

Condition data monitoring and aggregation techniques [4, 9] can be used to estimate the failure rates of most of the physical components of an NPP.

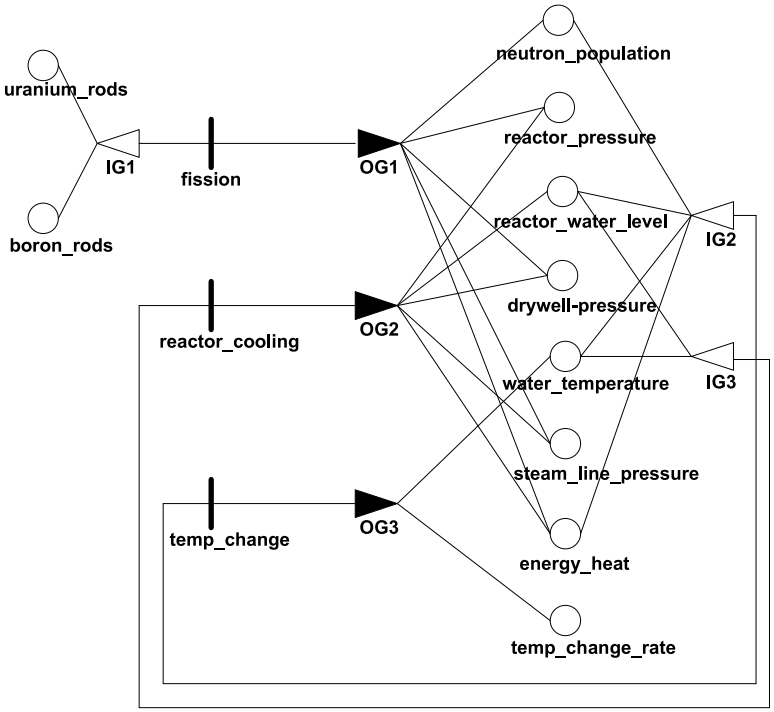


Figure 3. SAN model excerpt of NPP processes.

The time distribution functions of timed activities that model transients may be used to calculate the failure rates of physical components based on the numbers of tokens found at the moment of estimation in the SAN places used to model physical conditions. For example, the failure rate of feed_pump_trip in Figure 2 is computed based on the numbers of tokens in SAN places denoting various conditions of reactor feed pumps. The output functions of output gates associated with a timed activity that models a transient, in turn, models the perturbations to NPP operational variables by incrementing or decrementing the numbers of tokens in NPP places.

NPP operation modeling requires SAN models to be constructed for numerous reactor processes. Examples include nuclear fission, reactor cooling and the various physical relations between temperature, pressure, density, volume, etc. The rates of timed activities in these SAN models are provided in NPP technical manuals. The SAN models also incorporate timed activities that model NPP operation measures such as reactor period (time required for the reactor power to change by a factor of e, where e is the base of the natural logarithm) and reactivity (neutron population change with time). Figure 3 presents an excerpt of the SAN model developed for a set of reactor processes. The Join and Replicate operations of Möbius [6] are used to unify the atomic SAN mod-

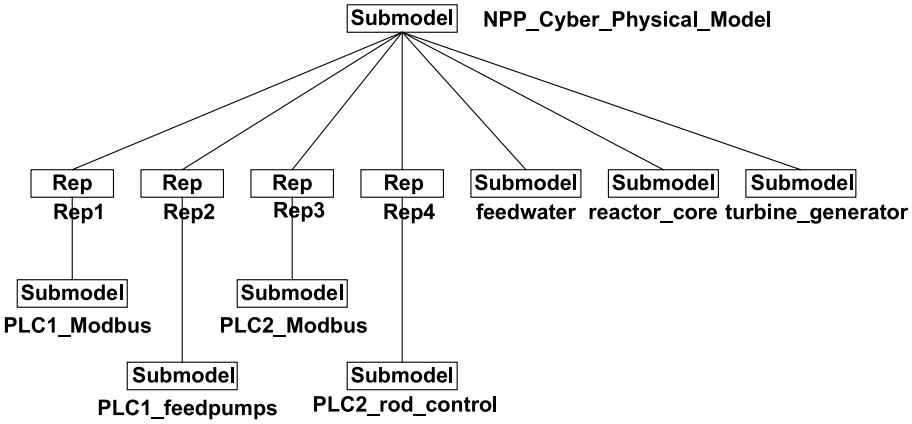


Figure 4. Composed SAN model excerpt for a boiling water reactor.

els developed for Modbus-based digital I&C systems, their effects on the NPP, and NPP processes.

Thus, a composed SAN model is constructed as a unified cyber-physical model of NPP operation. Figure 4 provides an excerpt of the composed SAN model developed for the boiling water reactor.

4. Estimating NPP Operation Profiles

Let P denote the set of device places defined in the atomic SAN models developed for the boiling water reactor. The set of possible markings of P is the finite set of functions $F = \{f_n \mid f_n : P \rightarrow N, n \epsilon N\}$. We assume that for each NPP power level l_m, where $m \epsilon N$, P is marked according to a function η from a finite (in most cases small) set of functions Γ_m, where $\Gamma_m \subset F$, and that Γ_m can be estimated. The set of functions Γ_m for a power level l_m depends on the actual values of NPP operational variables during the interval of time that an NPP has that state and the tasks (e.g., inserting or withdrawing control rods, changing the rate of water pumps, opening or closing valves) are carried out on NPP components by digital I&C systems.

NPP operational variables (e.g., nuclear fission data estimated by a neutron monitoring system) are stored as Modbus data items in the memory of digital I&C systems. On the modeling side, the data values are reflected by the numbers of tokens in the corresponding device places, which we call "conditions device places." Further, the tasks carried out by digital I&C systems on NPP components are initiated via a series of writes to the memory locations designated for the corresponding Modbus data items. For example, the rate of a reactor feed pump for 100% NPP power can be set by writing a value of (say) 84 in the holding register of a digital I&C system; for 75% NPP power, the value of the holding register may be set to 63, and so on.

The correspondence existing between values of Modbus data items in the memory of digital I&C systems, operations or tasks carried out by digital I&C systems on NPP components, and NPP operational variables allows profiles of normal NPP operations to be estimated. Thus, given normal values of NPP operational variables associated with a power level l_m, we estimate the possible values of discrete inputs, coils, input registers and holding registers, which may be considered as legitimate.

We also assume that when an NPP transitions from a power level l_m to a power level $l_{m\pm\lambda}$, where $\lambda \epsilon N$, the marking of P changes according to a finite set of marking transition flows, i.e., an ordered sequence of functions from Γ_m to $\Gamma_{m\pm\lambda}$. In general, throughout the operation of an NPP, a finite set of ordered sequences of functions from F exist that represent transitions of values of Modbus data items in digital I&C systems when the NPP is operated.

We leverage marking transition flows to construct NPP operation profiles. The normal operation of an NPP is simulated along with operational phenomena such as transients and the normal marking transition flows are estimated. Given an NPP at a power level l_m, the marking transition flow mechanism estimates if the semantics and payload of an arbitrary PDU received by a digital I&C system make sense in the Modbus configuration and will legitimately change the values of Modbus data items, respectively.

The set of functions Γ_m for each NPP power level l_m is constructed using reward models based on the activity-marking reward structure [13]. Each function in the finite set of functions F is assigned a rate of reward. A typical reward function for an element of F checks for a defined number of tokens in the conditions device places.

As mentioned above, NPP operational variables are reflected in Modbus data items in digital I&C systems that are modeled by conditions device places. The number of tokens in a SAN place represents a link between a power level l_m of an NPP and an element of F. In addition to these places, a reward function checks for the defined numbers of tokens in the remaining device places (i.e., device places that do not model NPP operational variables). In Möbius, the numbers of tokens in device places are parameterized using global variables. The composed SAN that models NPP operation is solved after all the global variables are assigned.

The following block of code implements a reward function:

```
double reward = 0;
if((PLC1_Modbus->holding_register_1->Mark()==reg_val_int1)&&
   (PLC1_Modbus->holding_register_2->Mark()==reg_val_int2) &&
   (PLC1_Modbus->holding_register_3->Mark()==reg_val_int3) &&
   (PLC1_Modbus->coil_1->Mark() == cval1) &&
   (PLC1_Modbus->discrete_input_1->Mark()==dval1) &&
   (PLC1_Modbus->input_register_1->Mark()==inp1))
{
   reward += 0.1;
}
return (reward);
```

The rates of reward fall in the interval-of-time category [13], where the total reward is accumulated during an interval of time covering all possible state transitions of an NPP. Solutions of the composed SAN model produce a reward for each element of F. The set of functions Γ_m for each power level l_m of an NPP consists of all the elements of F that map conditions device places to the numbers of tokens corresponding to values of NPP operational variables, in turn, identified by l_m, and whose accumulated reward is greater than 0.

Let S and V_m denote the set of conditions device places and the set of numbers of tokens that model values of NPP operational variables corresponding to power level l_m. Further, let R be a function that, for each element of F, returns the reward as estimated by composed SAN model solutions. Then, the set of functions Γ_m is $\{\mu \epsilon F \mid \mu : S \rightarrow V_m, R(\mu) > 0\}$.

Marking transition flows are identified by incremental measurement time intervals in Möbius. For instance, if solutions of the composed SAN model produce non-zero rewards of, say, $r_{4,8}$, $r_{1,2,0}$ and $r_{1,6,0}$ for the functions $f_{4,8}$, $f_{1,2,0}$ and $f_{1,6,0}$, respectively, solving the composed model again while setting the measurement time interval to a value t close to zero may produce zero rewards for these functions. Setting the measurement time interval to $t + 1$, $t + 2$, etc. and then solving the composed SAN model may produce a reward of $r_{1,2,0}$ for function $f_{1,2,0}$. Incrementing the measurement time interval again and solving the composed SAN model again may produce rewards of $r_{1,2,0}$ and $r_{1,6,0}$ for the functions $f_{1,2,0}$ and $f_{1,6,0}$, respectively.

The measurement time interval may be further incremented until the solution of the composed SAN model produces rewards of $r_{4,8}$, $r_{1,2,0}$ and $r_{1,6,0}$ for the functions $f_{4,8}$, $f_{1,2,0}$ and $f_{1,6,0}$, respectively, leading to the identification of the marking transition flow $f_{1,2,0}$, $f_{1,6,0}$ and $f_{4,8}$. The frequency of marking transitions caused by events such as transients is quite relevant to the attribution of possible losses of equipment during NPP operation. The estimation of the frequency of these marking transitions is carried out by defining impulse rewards associated with completions of timed activities that model NPP transients. Like the rates of reward, the impulse rewards defined in our work belong to the interval-of-time category of the activity-marking reward structure [13].

5. Deriving Intrusion Detection Rules

Our ultimate objective is to construct a set of practical intrusion detection rules. These rules scrutinize the values of Modbus data items and PDU fields to determine whether incoming PDUs are legitimate or malicious.

The SAN formalism is used to model regularities in Modbus data items and for marking transition flows so that the resulting SAN model is interpretable. The SAN model (Figure 5) is used to derive intrusion detection rules. The elements of P are modeled as SAN places. For each power level l_m of an NPP, the numbers of tokens in the SAN places are defined by elements of Γ_m. The activities in the SAN model capture individual memory writes carried out by Modbus requests.

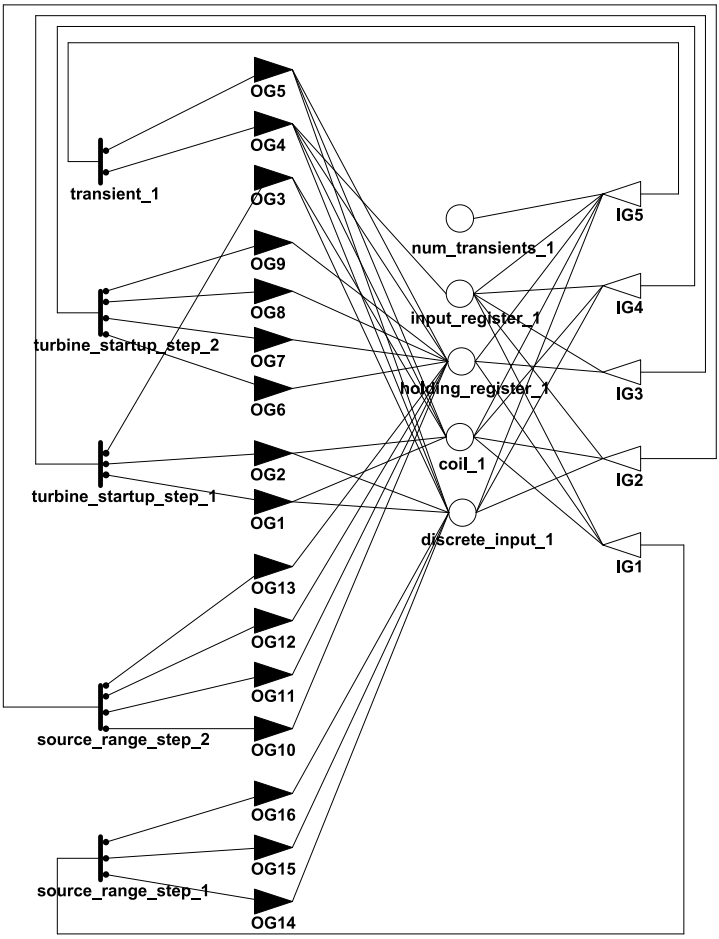
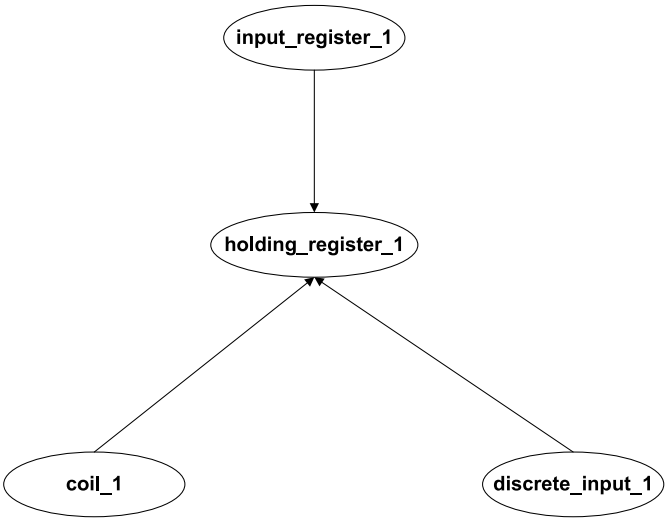


Figure 5. SAN model with data item regularities and transition flows.

For example, a PDU with function code `0x10` (write multiple registers) may write from 1 to 123 registers depending on the value in the registers field. Each of these memory writes is modeled by an activity in the SAN model. Thus, a marking transition is modeled as a set of activities. Activity cases model assignments of defined values to Modbus data items as a result of processing Modbus requests. The case distribution function of each activity models the probability distribution of a defined Modbus data item. Case distribution functions are constructed by analyzing the data provided by the set of functions Γ_m for each power level l_m, marking the transition flows and frequencies of marking transitions caused by transients.

These analyses are performed using the Bayesian belief network (BBN) formalism [10] as implemented by MSBNx [7]. Figure 6 illustrates the estimation of the probability distribution of a holding register using a BBN model. The



			holding_register_1			
input_register_1	coil_1	discrete_input_1	0x07f8	0x0be4	0x0c08	0x0ca8
0x0fd2	0	0	0.25	0.25	0.25	0.25
0x1f6c	0	1	0.75	0.0	0.25	0.0
0x03fc	1	1	1.0	0.0	0.0	0.0
0x02b6	1	0	0.0	0.5	0.5	0.0

Figure 6. BBN model estimating the probability distribution of a holding register.

probability is 0.75 that holding_register_1 is assigned the value 2040 (`0x07f8`) while heading toward a legitimate marking according to a legitimate marking transition flow. Activities in Figure 5 complete in order and the probability distributions of the Modbus data items change as the markings of the elements of P change. During NPP operations, Modbus data items and the PDUs received over the network are required to follow a SAN model as shown in Figure 5 in order to be considered as legitimate traffic in NPP operations. Monitoring and control traffic is examined using the intrusion detection rules derived from the SAN model.

The enabling predicates in each input gate check if the current value of a Modbus data item allows for a memory write that modifies a Modbus data item in compliance with a defined set of functions and following a legitimate marking transition flow. In fact, under normal NPP operations, conditions exist under which the defined valves are never opened or closed, control rods are never inserted or withdrawn, etc. No matter what the modifying value

is, the very action of modification may not comply with any NPP operation profile. Case distribution functions of activities and output functions (of output gates) are used to estimate the probabilities that legitimate values are assigned to Modbus data items. The following block of code consults the values of an input register, a coil and a discrete input, and returns a probability of 0.75 that 2040 (`0x07f8`) is assigned to the defined holding register in compliance with the operation profiles:

```
if((input_register_1->Mark() == 8044) && (coil_1->Mark()==0)
   && (discrete_input_1->Mark() == 1))
return(0.75);
holding_register_1->Mark() = 2040;
```

6. Conclusions

Operation-aware intrusion detection is a novel anomaly-based approach for detecting attacks on complex systems such as NPPs. Profiles of NPP operation as controlled and monitored by digital I&C systems are constructed using SAN models. The SAN formalism effectively models the interactions between a Modbus master and digital I&C systems, the operation of digital I&C systems on NPP components, and relevant NPP processes. The formalism also models and unifies NPP operation profiles estimated by the SAN model solutions. The resulting SAN model is interpretable as well as a valuable source of intrusion detection rules that can distinguish between legitimate protocol frames and attack frames.

Acknowledgements

The research of Julian Rrushi was partially supported by scholarships from the University of Milan and $(ISC)^2$.

References

[1] C. Bellettini and J. Rrushi, Vulnerability analysis of SCADA protocol binaries through detection of memory access taintedness, *Proceedings of the IEEE SMC Information Assurance and Security Workshop*, pp. 341–348, 2007.

[2] D. Deavours, G. Clark, T. Courtney, D. Daly, S. Derisavi, J. Doyle, W. Sanders and P. Webster, The Möbius framework and its implementation, *IEEE Transactions of Software Engineering*, vol. 20(10), pp. 956–969, 2002.

[3] R. Krutz, *Securing SCADA Systems*, Wiley, Indianapolis, Indiana, 2006.

[4] J. McCalley, Y. Jiang, V. Honavar, J. Pathak, M. Kezunovic, S. Natti, C. Singh and J. Panida, Automated Integration of Condition Monitoring with an Optimized Maintenance Scheduler for Circuit Breakers and Power Transformers, Final Project Report, Department of Computer Science, Iowa State University, Ames, Iowa, 2006.

[5] J. Meyer, A. Movaghar and W. Sanders, Stochastic activity networks: Structure, behavior and application, *Proceedings of the International Conference on Timed Petri Nets*, pp. 106–115, 1985.

[6] J. Meyer and W. Sanders, Specification and construction of performability models, *Proceedings of the Second International Workshop on Performability Modeling of Computer and Communication Systems*, 1993.

[7] Microsoft Research, MSBNx: Bayesian Network Editor and Tool Kit, Microsoft Corporation, Redmond, Washington (research.microsoft.com /adapt/MSBNx).

[8] Modbus IDA, MODBUS Application Protocol Specification v1.1a, North Grafton, Massachusetts (www.modbus.org/specs.php), 2004.

[9] J. Pathak, Y. Jiang, V. Honavar and J. McCalley, Condition data aggregation with application to failure rate calculation of power transformers, *Proceedings of the Thirty-Ninth Annual Hawaii International Conference on System Sciences*, p. 241a, 2005.

[10] J. Pearl, Bayesian networks: A model of self-activated memory for evidential reasoning, *Proceedings of the Seventh Conference of the Cognitive Science Society*, pp. 329–334, 1985.

[11] W. Sanders, Construction and Solution of Performability Models Based on Stochastic Activity Networks, Ph.D. Dissertation, Department of Electrical Engineering and Computer Science, University of Michigan, Ann Arbor, Michigan, 1988.

[12] W. Sanders, Integrated frameworks for multi-level and multi-formalism modeling, *Proceedings of the Eighth International Workshop on Petri Nets and Performance Models*, pp. 2–9, 1999.

[13] W. Sanders and J. Meyer, A unified approach for specifying measures of performance, dependability and performability, in *Dependable Computing for Critical Applications*, A. Avizienis and J. Laprie (Eds.), Springer-Verlag, Berlin-Heidelberg, Germany, pp. 215–237, 1991.

[14] W. Sanders and J. Meyer, Stochastic activity networks: Formal definitions and concepts, in *Lecture Notes in Computer Science, Volume 2090*, Springer, Berlin-Heidelberg, Germany, pp. 315–343, 2001.