Detecting Sensors and Inferring their Relations at Level-0 in Industrial Cyber-Physical Systems

Kalyan Perumalla*, Srikanth Yoginath*, Juan Lopez†
* Computer Science and Mathematics Division,
† Cyber and Data Analytics Division,
Oak Ridge National Laboratory,
Oak Ridge, TN USA
Email: perumallaks,yoginathsb,lopezj@ornl.gov

***Abstract*—While there exist tools and techniques to discover, identify, map, and analyze cyber physical components at higher levels of the cyber space, there is a lack of capabilities to reach down to the sensors at the bottom-most levels, such as levels 0 and 1 of the Purdue Enterprise Reference Model for cyber-physical systems (CPS). Conventional information technology (IT)-based tools reach as far as the network-side of programmable logic controllers, but are inadequate to access and analyze the physical side of the CPS infrastructure that directly interfaces with the actual physical processes and systems. In this paper, we present our research and development efforts aimed at addressing this gap, by building a system called DEEP-CYBERIA (Deep Cyber-Physical System Interrogation and Analysis) that incorporates algorithms and interfaces aimed at uncovering sensors and computing estimates of correlations among them.**

***Index Terms*—sensors, correlations, machine learning, deep learning, causality, programmable logic controllers, inference,**

I. INTRODUCTION

In this paper, we identify a gap in capabilities in mapping sensors of cyber-physical systems, and, to bridge the gap, present our ongoing efforts in progressing the capabilities from no knowledge to increasing levels of knowledge about sensors at the bottom-most layers of a cyber-physical asset. Sensors form a critical layer of many cyber-physical systems such as water treatment plants, electric grids, and nuclear reactors. In gaining an accurate and deep view of a cyber-physical system, the challenge is to see beyond the network side of the system to the side where sensors are connected. In the layered architecture of cyber-physical system, relatively few technologies exist to discover the bottom-most level, namely, level-0. Level-0 is made of the sensors connecting the cyber components of the cyber-physical system (CPS) to its physical components. Our approach to addressing this goal is to develop a new sensor data assimilation and inference engine called DEEP-CYBERIA. Here, we present the motivating factors behind this level-0 inference system, describe the technical approach, and present case study-based experimental results to showcase the current capabilities of DEEP-CYBERIA. We also outline additional ongoing and future work, especially focused on generalizing the approach, software and hardware test-beds to address the inference problem across multiple other domains, ultimately aimed at achieving the goals in a domain-agnostic fashion.

This manuscript has been authored by UT-Battelle, LLC under Contract No. DE-AC05-00OR22725 with the U.S. Department of Energy. The United States Government retains and the publisher, by accepting the article for publication, acknowledges that the United States Government retains a non-exclusive, paid-up, irrevocable, worldwide license to publish or reproduce the published form of this manuscript, or allow others to do so, for United States Government purposes. The Department of Energy will provide public access to these results of federally sponsored research in accordance with the DOE Public Access Plan (http://energy.gov/downloads/doe-public-access-plan).

A. Background

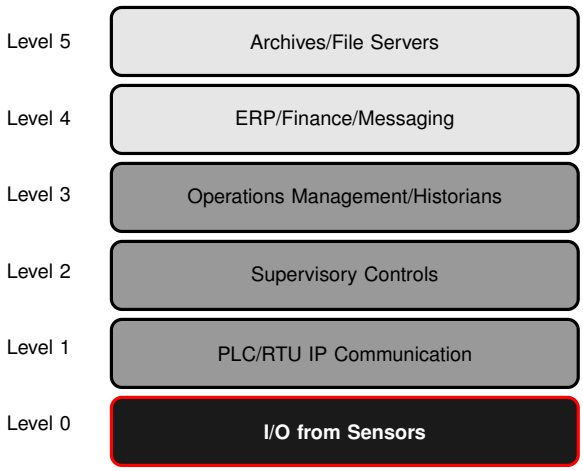

Fig. 1. The layers of industrial cyber-physical systems based on the Purdue Reference Architecture Model

Figure 1 shows the Purdue Enterprise Reference Architecture Model (PERA) [1] in which various aspects of a cyber-physical system are organized according as layers of functionality. The bottom-most layer is that which directly interfaces the rest of the cyber infrastructure to the controlled features and sensed phenomena of the physical processes in the cyber-physical system. In large, complex, and long-lived installations, these sensors are designed to be accessed over various controller protocols and are not directly accessible or addressable. Probing, detecting, inferring, or otherwise manipulating them individually is a difficult problem, especially in a portable, system-oblivious manner.

Note that this PERA model is specially designed for Operational Technology (OT) networks of CPS across many industries. This PERA model is distinct from the Information Technology (IT) networks that are specific to the IT sector that have entirely different applications and scope.

B. Motivation

There are several reasons why discovery, identification and analysis of sensors is of great interest. Figure 2 illustrates some of the salient reasons.

The following are some of the challenges associated with sensors at level-0 of the PERA model.

- Lost documentation: Due to the age of the CPS, precise documentation about the components, their interconnections, sensor types, etc. are either not maintained or lost in the organization. Owing to the long operating lives of the sensors, they may still be perfectly functional, and may be expected to serve long after their documentation is lost.
- Diagnosis: A malfunctioning CPS may need to be analyzed with precise information about actual sensor types and their behaviors.
- Verification and audit: Although blueprints or other documentation may be available, the ground truth would need to be independently discovered in order to serve as input to independent audit and verification procedures.
- Detection of misconfiguration: Comparison of intended versus actual behaviors needs to be performed on an actual operating CPS. To help in the comparison, sensors would need to be discovered and analyzed.
- Hard to access locations: Some CPS installations present difficulties in taking stock or inventorying the sensor installations, requiring remote, automated discovery and verification.
- Triage: When incidents such as mishaps or natural disasters occur, documented information is largely inadequate and unreliable in taking stock of the status. Automated discovery and identification would be necessary.
- Intelligence gathering: In addition to normal, civilian use cases, there are many applications in the intelligence space that rely on sensor discovery.

C. Approach

In the context of the preceding set of motivation and challenges, we are addressing the sensor mapping problem by developing a new system called DEEP-CYBERIA. Given a cyber-physical system, DEEP-CYBERIA tackles the sensor mapping problem as follows. Just as blood samples are drawn and analyzed to diagnose a human body's constitution and health, we draw samples of network traffic from the CPS and perform intelligent analysis to discover its sensors and their correlations. Initial intelligence is obtained using passive techniques. Future work involves application of active techniques to carefully probe and interact with the CPS for gleaning enhanced knowledge.

DEEP-CYBERIA has the capability to ingest packet capture (`pcap`) data or CPS historian data. DEEP-CYBERIA's analysis capabilities are currently tested with the `Modbus` protocol and can be easily extended to other CPS protocols. For any given target CPS, we iterate through the input data with different data analyses procedures to uncover target CPS specific characteristics. Some of the current capabilities are listed below.

II. DISCOVERY AND ANALYSIS

Exploratory Indicators and Metrics

To aid in uncovering the underlying relations, we have developed and incorporated multiple indicators and metrics of the inter-relations among the gleaned sensor streams. Our ideal goal is to be able to deal with any CPS, without requiring deep knowledge of its components, physical processes, physical units, etc. This implies that we would need to glean and generate categories and statistical measures in a general way. The following are some of the explorator indicators and metrics we have incorporated in DEEP-CYBERIA.

- **Data Descriptors**: We categorize the time-series data from the sensors obtained from the CPS into different categories such as *constants*, *binary*, *ordinal* and, *continuous*. This categorization allows us to broadly relate input data streams in terms of their functionality and also helps in filtering input data streams. For example, *constant* data-streams are not needed for analyses that determine causal relations. Additional pre-processing such as smoothing can also be performed on the raw data.
- **Auto-Correlation**: In a continuously running CPS, we intend to see repeating patterns in the input data streams. Generally, these patterns reveal the periodicity of a given data stream and their periodic correspondence to other data streams.
- **Cross-Correlations**: Pearson's cross-correlation reveals the correlation between the data streams, which might be helpful in relating certain cases. For example: voltage on certain component like, valve can have direct correspondence with the liquid level in the container. Hence, a strong positive or negative correlation are usually related and can be binned together to a sub-system or a container in our example. We also use Kendall-Tau correlation, in which the concordant and discordant pairs actually determine the correlation between the any two input data streams. This is helpful statistic to establish ordinal association between two input data streams.

Dependency Graphs

Based on the domain-agnostic, exploratory indicators and metrics automatically generated from the sensor data streams, the next level of processing involves the reconstruction of possible dependencies among the sensors *as they originally exist in the physical components and manifest in the cyber components*. The following are some of the relations thus generated, in the form of graphs of dependencies (causal or correlated in nature) among the sensors.

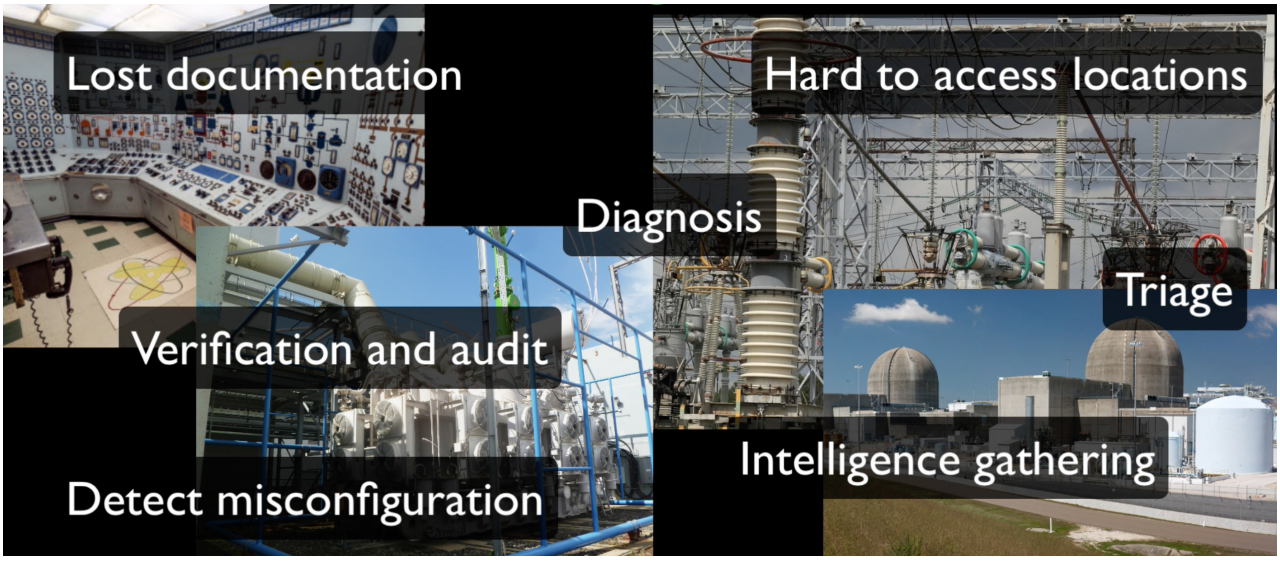

Fig. 2. Motivating reasons behind discovery, identification and correlation-analysis of sensors in complex CPS installations

- **Granger Causality**: This metric is used to generate directed causal graphs capturing the causality among sensors, where individual nodes are time-series data streams from the input. The core principle of behind Granger causal relation [2] is that any data stream X causally affects Y, if Y can be predicted above and beyond just by using previous values of X. When we execute the Granger causality procedure on all input data streams picking two streams at once, we obtain a causal graph.
- **Heuristic-based Causal Graphs**: In this method we look for specific events that result in drastic change in data readings. For example: a binary signal changes its state from 0 to 1 or a more concrete example would be when voltage value shoots from its minimal to its maximum value to activate an actuator like, a valve. From this point in time, we look at the changes that every input data stream undergoes before and after this event. If we detect such changes in any input data stream, we consider the detected event has a causal affect on that input data stream. We than generate a graph based on such information to obtain a causal graph.

Classification

In the classification module, we perform a training process on a data model to classify the byte stream extracted directly from the series of packet payload sequences into certain categories. Here, the signatures of various data-streams as they occur within the packets are learned. Anomalies can be marked when the classifications predicted from the trained model consistently and drastically differ from expectations. Two supervised machine learning models are being used in DEEP-CYBERIA, one based on random-forest and the other based on convolutional neural networks (CNN).

III. CASE STUDIES: ANALYSIS IN ACTION

We have applied the statistical procedures listed in Section II on time-series data from CPS of different scale. Some are of lab-scale like, the simplistic water-cascade CPS system that is configured to emulate operations of real-life CPS systems, while others are of campus scale like High-Flux Isotope Reactor. As a case study, to demonstrate the efficacy of the DEEP-CYBERIA statistical toolset, we use the canal-lock CPS emulation realized using our lab-scale water-cascade CPS.

A. A CPS Testbed for DEEP-CYBERIA

Figure 3 illustrates a physical canal lock in which a ship moves from a lower level to a higher level of water body by first entering to chamber1 through gate1. Chamber1's water level is increased to some intermediary level for the ship to cross over to chamber2 through gate2. Once the ship is in chamber 2, the water level in this chamber is increased from the intermediary level to the higher water body level. At this point, the ship exits the chamber3 and the Canal Lock CPS through gate3. A similar sequence of steps in the reverse-order is followed when the ship needs to travel from the higher-level to the lower level of water body.

The emulation of Canal Lock CPS was realized by using four acrylic tanks, labeled T1, T2, T3, and T4, as illustrated in Figure 3. The first tank T1 and the last tank T4 serve as lower-level and upper-level water bodies of the physical Canal Lock CPS, respectively. Tanks T2 and T3 act as the chamber1 and chamber2 of the Canal Lock CPS. Tanks T2 and T3 are equipped with pumps P2 and P3, and valves V2 and V3, respectively. The pumps are immersed in a reservoir, which holds water to fill all the tanks. Valves V2 and V3 drain water directly into the reservoir. The Allen-Bradley Micrologix 1100 PLC with two extension slots are used to maintain, control and emulate the Canal Lock CPS operational behavior.

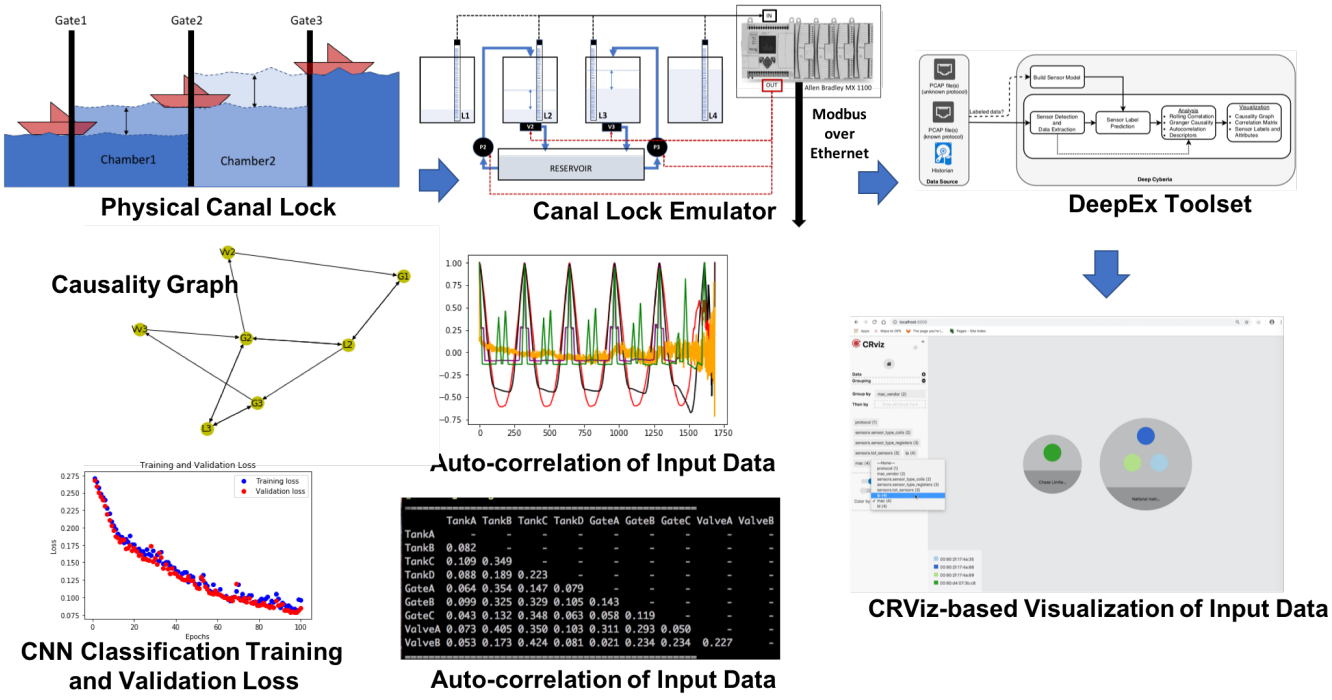

Fig. 3. DEEP-CYBERIA case study using a water cascade CPS to emulate physical canal lock behavior. The Ethernet packets are captured from live-network, filtered for Modbus protocol payload and used as input for our DEEP-CYBERIA analyses tools. The data-descriptor, auto-correlation, cross-correlation, deep-learning training and validation and causal graph outputs are shown.

For the analyses the water cascade system was configured to continuously emulate the UP scenario and Down scenario, where the ship in Figure 3 moves from left-to-right and right-to-left, respectively. During this period the input data was collected. The input data for our analyses were the packet captures of the Modbus traffic. These captures are directly fed to our DEEP-CYBERIA analyses routines.

B. DEEP-CYBERIA *Analyses in Action*

Data Descriptors: Each data stream from the input are identified into categories specified in Section II. This data is visualized using the CRviz [3] tool, which provides highly flexible viewpoints of the CPS data-sets. One such example is from the PLC point of view, which displays different PLCs, whose components can be viewed by clicking on the corresponding icon, the interface itself is exploratory by design. We also have integrated this with the database NIC, PLC and other vendors, which are detected and populated in CRviz visualization tool.

Auto-correlation: The auto-correlation graph in Figure 3 show the existence of periodic cycles in the target CPS. The *red* and *black* lines in the auto-correlation figure correspond to the level-sensor values of the tanks T2 and T3, respectively. As seen the signal pattern of both these level-sensors consistently repeat because of continuous emulation of UP and DOWN scenarios. With this graph, we can definitely say that the CPS is a periodic process and repeats after every 350 seconds.

Cross-correlation: We compute cross-correlation values across different input data-streams. The relations between voltages on valves and the level-sensor values have relatively strong correlation. The valve voltage is usually at its minimum value but shoots up suddenly and reduces slowly with its rate of decrease almost equivalent to the out-flow of water. Hence, there exists a strong correlation between the valve voltages and the level-sensor values. When provided with data-stream with two signals having strong correlation with each other, we might be able to associate them to be part of one sub-system, like we associated the level-sensor and valve-voltage to a Tank in this case study.

Classification: The training and validation errors as the model is trained with input data from the network is shown in Figure 3. The abscissa shows the number of epochs, which is the number of times the input is reexamined while learning. The ordinate is the percentage of the input for which the model incorrectly predicts the label. We used the trained model to detect the categories of the Modbus payloads extracted from new packet captures and used these labels for further analyses.

Causality Graphs: The data comprised three binary signals namely, the statuses of Gates (G1, G2 and G3), they are 0 when the gate is closed and 1, when the gate is open. The causality graph was constructed using the transition of these signals as events. The causality graph shown in Figure 3 correctly identifies that gate ”G2” the center one affects the valve voltages and water levels of tanks T2 and T3. Similarly, causal relations shown by all other relations can be correctly reasoned. With this graph, we are able to identify the association between different data points in the Modbus

payload of an Ethernet packet capture.

IV. Generalizing the Test-bed

A major challenge in achieving the goals of inference is to carry the algorithms from one system to another. However, it is extremely difficult, time-consuming, and expensive to build a surrogate system for every type of application. The test-beds described in Section III were assembled with a significant amount of effort and domain-level expertise; from this exercise, it became clear that building such test-beds for the purpose of creating the inference system is nearly impractical for every target CPS domain or application. For example, when we tried to retarget and retrain DEEP-CYBERIA on a subsystem of a nuclear isotope generation facility at our lab, it became clear that it is impractical to assemble a surrogate hardware test-bed that replicates both physical and cyber parts of that facility, just for the purposes of training and validating our inference algorithms and engine.

It became clear that it is highly desirable to have a test-bed that can be configured to serve as a surrogate for *any* CPS such that it can drive the development of DEEP-CYBERIA. Instead of building a new hardware prototype for each case, it is effective to have a hardware test-bed that can accept sensor inputs that are extracted from any desired installation and, using those inputs, drive a network of actual PLCs that will behave almost exactly as though the PLCs were installed in the original system of interest.

In essence, the idea of this generalized test-bed is as follows. The cyber component of the CPS of interest is conceptually separated from the physical component of that CPS by disentangling them at the level 0-1, which is the actual sensor device readings that are perceived on the wires by the PLCs' sensor input ports. It is relatively easy to duplicate the PLC and communication network part of the CPS than duplicating the physical part. However, the physical part is indistinguishable from the streams of sensor input values themselves that are received by the PLCs. Therefore, if we can "cut" the system at that boundary, we would essentially have a way to instantiate a generalized test-bed that has reusable and retargetable cyber component. This cyber component would then consist of custom-driven sensor drivers, actual PLCs, actual networks, actual data historians and actual analyzers and controllers. This offers the best of both worlds: the decoupling of the physical component from its sensor observations while not losing the high fidelity of cyber component needed for accurate analysis.

A generalized test-bed of this kind also offers additional, novel benefits: it makes it possible to perform repeatable experiments in which the physical component behaviors can be repeatedly studied across multiple runs, making it possible to debug, test, refine, and enhance the sensor inference systems such as our DEEP-CYBERIA system.

We are currently in the process of assembling such a generalized test-bed system. The hardware for this test-bed includes (a) a range of representative PLCs, (b) daughter boards to convert any given sensor value to a signal on the inputs of the PLCs, (c) Raspberry PI devices to drive their attached daughter boards, and (d) software to convert historian data of sensor value streams into accurately timed, translated, converted, and triggered inputs at the PLC inputs via the daughter boards.

V. Summary and Future Work

In advancing the state-of-the-art in CPS intelligence, DEEP-CYBERIA is moving beyond traditional cyber-surface interrogation to deep sensor interrogation and beyond. DEEP-CYBERIA focuses on the development of a broadly applicable, novel capability to deepen understanding of a target cyber-physical assets. Starting with the data from the network packet capture files or data dumps from CPS historians, DEEP-CYBERIA strives to illuminate the topological, functional and behavioral specifics of a target CPS.

We have developed and incorporated multiple statistical, machine-learning and domain-specific heuristic-based tools into DEEP-CYBERIA to help in this direction. Our efforts are aimed at developing a network discovery capability (both passive and active) to enhance discovering, monitoring, and diagnosing the identity of cyber-physical system (CPS) components. The interrogation and analysis capabilities are intended to advance the state-of-the-art by going beyond the internet/network-level probing and inference by inferring sensor elements behind network-addressable controllers at the level-0 in the Purdue Model of Control Hierarchy [4].

Based on discovered sensor elements, analysis capabilities are targeted to uncover inter-dependencies among sensors with respect to cyber and physical process interactions, triggers, and after-effects. Analysis capabilities are aimed at building the foundation for sophisticated forensic features that reach beyond basic data-based inference. In addition to small CPS test-beds, as a complex case study, the experimental network of the Cold Source portion of the High Flux Isotope Reactor (HFIR) facility at ORNL is being exercised with the DEEP-CYBERIA implementation.

Acknowledgements

Research was sponsored by the Laboratory Directed Research and Development Program of Oak Ridge National Laboratory, managed by UT-Battelle, LLC, for the US Department of Energy under contract DE-AC05-00OR22725.

References

[1] Wikipedia, "Purdue enterprise reference architecture," https://en.wikipedia.org/wiki/Purdue_Enterprise_Reference_Architecture, 2019.

[2] C. W. Granger, "Causality, cointegration, and control," *Journal of Economic Dynamics and Control*, vol. 12, no. 2-3, pp. 551–559, 1988.

[3] B. Andrea, "Crviz: Scalable design for network visualization," https://www.cyberreboot.org/projects/crviz/, 2018.

[4] T. J. Williams, "The purdue enterprise reference architecture," *Computers in Industry*, vol. 24, no. 2, pp. 141 – 158, 1994.