A New Approach to Insider Threat Detection & Mitigation for Critical Infrastructure

Artificial Neural Networks and Risk Significance

Adam D. Williams*, Shannon N. Abbott, Chris A. Faucett, Colton Heffington, Sondra Spence

Sandia National Laboratories, adwilli@sandia.gov

William Charlton

Nuclear Engineering Teaching Laboratory, University of Texas

Katherine Holt

Office of International Nuclear Security, U.S. National Nuclear Security Administration

To improve insider threat detection and mitigation (ITDM) for critical infrastructure, some efforts—including the U.S. Department of Homeland Security's Cyber and Infrastructure Security Agency 2020 Insider Threat Mitigation Guide—have shifted focus toward evaluating "the potential for an insider…to harm that organization." This shift is an opportunity for a paradigmatic evolution in traditional ITDM programs that emphasize preventative and protective strategies. In response, a joint study by Sandia National Laboratories and the University of Texas' Nuclear Engineering Teaching Laboratory explored a data-driven, collective behavior-based approach to systemically model insider potential. This new approach leverages advances in artificial neural networks, invokes tenets from organization science, and incorporates risk significance concepts. As a result, this research demonstrates the benefit of using empirically derived operational patterns and workplace rhythms as baselines of expected behaviors from which to detect anomalies—captured in terms of specific types, frequencies, magnitudes, or quantities of deviations—for improving ITDM in critical infrastructure.

CCS CONCEPTS • **Security and privacy → Human and societal aspects of security and privacy; • Computer systems organization →** *Embedded and cyber-physical systems; Sensor networks*;

Additional Keywords and Phrases: Insider threat, artificial neural networks, detection & mitigation, risk significance

ACM Reference Format:

Adam D. Williams, Shannon N. Abbott, Chris A. Faucett, Colton Heffington, Sondra Spence, William S. Charlton, and Katherine Holt. 2023. A New Approach to Insider Threat Detection & Mitigation for Critical Infrastructure: Artificial Neural Networks and Risk Significance. WRIT Proceedings Paper TBD, TBD, Article TBD (TBD 2023), TBD pages.

1 INTRODUCTION

Recent trends have challenged traditional approaches to insider threat detection and mitigation (ITDM) for critical infrastructure. Legacy insider threat mitigation programs tend to focus on identifying characteristics of individuals related

*Corresponding author.

to possible malicious actions. More specifically, consider the U.S. National Insider Threat Task Force (NITTF) definition for insider threat:

> *the risk [that] an insider will use their authorized access, wittingly or unwittingly, to do harm to their organization. This can include theft of proprietary information and technology; damage to company facilities, systems or equipment; actual or threatened harm to employees; or other actions that would prevent the company from carrying out its normal business practices. [1, p. 3]*

In response, many insider threat mitigation approaches in critical infrastructure tend to emphasize *preventative* (measures implemented before access is granted) and *protective* (measures taken after access is granted and throughout employment) strategies to mitigate unwanted individual behaviors [2]. Such approaches can easily manifest an overreliance on generic job task analysis and detection of aberrant individual behavior that may not fully account for workplace behavior patterns or sufficiently capture facility recovery operations.

Yet, there is a benefit from shifting away from this traditional focus toward a new conceptual paradigm for contextualizing and addressing insider threats. For example, the 2011 U.S. Executive Order that created the NITTF launched a coordinated effort to establish insider detection and prevention programs [3]. In addition, its 2017 best practices publication asserted "a dynamic effort requiring constant evaluation, fresh perspectives, and updated approaches [4]" to vigilantly, diligently and successfully mitigate insider threats in critical infrastructure. Likewise, consider how the U.S. Department of Homeland Security's Cyber and Infrastructure Security Agency (DHS/CISA) defined inside threat as "the *potential* for an insider to use access or special understanding of an organization to harm that organization" (emphasis added) in their 2020 Insider Threat Mitigation Guide [5]. Transitioning away from common preventative and protective strategies to re-evaluate opportunities for exploring "the potential for an insider" to act offers new analytic opportunities.

One such opportunity centers on invoking collective behaviors observed in the workplace toward a more comprehensive "health-monitoring" approach to detecting and monitoring for insider threats. More specifically, observed—and empirically measured—patterns of expected operational activities can serve as a baseline from which to detect potential insider threat activities. From this perspective ITDM approaches seek to identify undesired deviations from expected (or normal) patterns of organizational behavior, as they *may* indicate an increased likelihood or opportunity for a malicious insider act. Simply put, if anomalies in expected operational patterns are identified, then opportunities for a successful insider act may also be identified. Yet, the allure of a collective behavior-based approach to ITDM is challenged by the need to distinguish clearly between natural organizational evolution and malicious intent in such anomalies.

Recent collaborative research efforts between Sandia National Laboratories and the University of Texas' Nuclear Engineering Teaching Laboratory (NETL)—sponsored by the National Nuclear Security Administration's Office of International Nuclear Security (NNSA/INS)—have explored the efficacy of a new technical and conceptual approach to ITDM. Invoking artificial neural networks (ANN) and the concept of risk significance, this research has demonstrated the ability for a collective-behavior-based ITDM approach to evaluate observable measures and organization-level indicators to better understand, detect, analyze, and mitigate insider threats. More specifically, the ability for ANNs to be "trained" to learn operational workplace patterns, and alert upon certain types, frequencies, or quantities of deviations, was explored. Similarly (and borrowing from the field of nuclear safety), anomalies in expected operational patterns and potential insider acts were conceptualized in terms of risk significance to enhance this empirical and data-driven ITDM approach.

After briefly situating this collective behavior-based approach for ITDM within relevant literature, this article next provides a longer description of ANNs, elements of organization science, and risk significance as the technical and conceptual basis for a new ITDM framework. Following a summary of the data collection and experimental design for the research is an exploration of the analytical results related to detecting insider threat deviations from expected operational workplace patterns. Lastly, this article discusses conclusions, insights, and implications from this data-driven and collective behavior-based ITDM approach.

2 SITUATING A NEW APPROACH TO INSIDER THREAT DETECTION & MITIGATION

While efforts to mitigate insider threat may be tailored to a specific type of critical infrastructure, many of the fundamental assumptions and much of the underlying philosophy is similar. For clarity, consider the internationally accepted approach for addressing insider threat in commercial nuclear facilities. In general, this approach focuses ITDM efforts on countering "an individual with authorized access to [nuclear material, associated facilities or … information … who could commit, or facilitate the commission of criminal or intentional unauthorized acts [...] directed at nuclear material … or other acts determined by the State to have an adverse impact on nuclear security" [6].

One of the core assumptions of this approach is that insider threat opportunities manifest given the right combination of access, authority, and knowledge that a given individual may have of a specific nuclear facility [7]. By extension, this approach assumes that threat opportunities materialize into attacks when an insider is motivated to act maliciously. The traditional set of responses revolve around preventive (e.g., efforts typically implemented *before* access to a nuclear facility is granted) and protective (e.g., efforts typically implemented *after* access is granted to a nuclear facility) measures to counter such insider acts. Simply put, preventive measures aim to reduce the likelihood of initially gaining opportunities to act maliciously and protective measures aim to reduce any gained opportunities to manifest into malicious acts. Examples of preventive measures include pre-employment screening, human reliability programs (HRPs), and other behavioral reporting mechanisms, where protective measures can include access controls, contraband detection, and other physical or cyber security measures. In the commercial nuclear domain, this approach to addressing insider threat is globally accepted and supported by both the International Atomic Energy Agency (IAEA) [6] and the World Institute for Nuclear Security [8] (WINS)—which are also consistent with the broader NITTF [1,4] definition previously discussed.

2.1 A New Data Approach to ITDM: Operational Patterns & Workplace Rhythms

To date, the collective behaviors of personnel within critical infrastructure have been an underutilized data set for developing ITDM approaches. While individual human behaviors can be extremely challenging to predict or anticipate, humans working in organizations tend to exhibit a set of operational patterns or workplace rhythms to execute their regular job duties. The ability to capture and evaluate data describing such patterns and rhythms could form the foundation of new ITDM approaches that use observable patterns of expected *collective* behaviors to overcome challenges in accurately identifying possible *individual* motivations for insider actions. Consider, for example, two popular concepts from the discipline of organization science useful for such a framing of ITDM. The first argues that behaviors within organizations can evolve from plans to implementation—suggesting a need to differentiate (and reconcile) between "as designed" and "as built" [9] conceptions. The second asserts that behaviors within organizations result from recurrent *individual* human action that is both (and simultaneously) shaped by artifacts and constructed by their *collective* interpretation [10]. Taken together, these two concepts suggest that ITDM solutions can be improved by incorporating the differences between planned insider threat mitigations ("as designed") and daily work practices with those mitigations ("as built"), as well as interpreting collective patterns in access, authority, and knowledge.

As described in more depth in [11], the relationships between individual actions, collective behaviors, and organization performance are dynamic and interdependent. For example, as individuals tend toward "short cuts" to efficiently execute regular job duties, organizational responses are prompted to promote desired levels of quality and consistency. In other words, if individual actions tend toward the limit of organizationally acceptable behaviors, then organizations will need to actively (and regularly) reinforce boundaries of functionally acceptable behavior. Consider, for example, diffusion of particles in a liquid, where the apparently random movements of individual particles give rise to an analyzable rate of spread through the liquid contained in a tank. Similarly, the constant (re)balancing of individual "Brownian movements" (like the particles in the liquid) and organizational "counter gradients" (the surfaces of the tank containing the liquid) reflect dynamics that can be captured by monitoring relevant, facility-level data signals over time and identifying natural operational patterns. Establishing sets of baseline patterns from sets of continuously collected facility-level data signals can help illustrate "error margins" between perceived and actual boundaries of organizationally acceptable behavior—including manifesting as thresholds for determining undesired deviations in such patterns and rhythms. By extension, this manner in which organization science uses observed patterns of workplace behaviors to understand (and monitor for) tendencies of individuals toward the limits of functionally acceptable behaviors offers a new framing for ITDM. In short, a better understanding of organizational patterns could improve ITDM programs by defining—and measuring—"the *potential* for an insider to … harm that organization" (emphasis added) [5] in terms of deviations from expected operational workplace patterns.

2.2 A New Technical Approach to ITDM: Artificial Neural Networks

The extent to which facility-level data signals can capture observable patterns in everyday workplace behaviors is predicated on the ability to effectively and comprehensively evaluate these data to elicit potential deviations in anticipated individual patterns. Advances in machine learning offer a set of options that can support the data-driven need to describe—and identify anomalies within—observed operational patters and workplace rhythms. Machine learning methods are often considered "black boxes" with limited transparency for explanation or interpretation—but are generally a (set of) algorithms capable of performing tasks without explicit programming. For example, machine learning methods can aggregate multiple, disparate data signals and detect anomalies defined as deviations from an expected baseline. While direct quantification relating to ITDM is likely to be difficult (or perhaps not even beneficial), invoking machine learning methods can significantly support a data-driven, collective behavior-based ITDM approach.

Despite the range of potential of relevant machine learning methods, artificial neural networks (ANN)—inspired by modeling neurons in a biological brain—provides useful characteristics for this new ITDM approach. At a high level, ANNs model neurons by organizing them into layers that form a network, where neurons are able to receive multiple input signals and produce (sets of) output(s). From this general ANN perspective, neuron signals across various layers are summed together, conditioned with a bias term, and applied within an activation function to ensure that the ANN can learn a nonlinear function. Generated output signals are propagated as inputs into other neurons throughout the network until a final output is calculated. The difference between the calculated and anticipated output value is back-propagated through the network for adjustment to help drive future predictions that are closer to true values—a generic process called "training." Iterating on the generic process, ANNs would be able to theoretically learn a range of potential (non)linear functions necessary for highlighting anomalies in anticipated operational patterns.

More specifically for a collective behavior-based ITDM approach, ANNs have been applied successfully to such domains as pattern detection, routine task performance reduction, and sensor attack mitigation. Relevant data streams that could potentially produce ITDM-relevant patterns could exhibit complex temporal or spatial dependencies in high dimensional data evaluation that challenge the applicability of traditional statistical approaches for detecting anomalies. ANN-based approaches, on the other hand, can help overcome these challenges and enhance the ability to capture subtle changes within larger, multisensory datasets related to anomalies. Though not a panacea, incorporating ANNs into ITDM solutions is likely beneficial. Yet, successful incorporation would need to address the infrastructure need for large amounts of training data, the potential difficulties in transferring algorithms between critical infrastructure facilities, calibrating for background signal noise, correcting for sensor drift (or misalignment), and protecting the ANN algorithms themselves from attack. While both the process and the potential limitations are discussed in more depth in [11], ANNs provide an additional tool—and set of analytic insights—to enhance a data driven, collective behavior-based approach to ITDM.

2.3 A New Conceptual Approach to ITDM: Risk Significance

Taken together, an ITDM approach centered on capturing deviations from expected operational workplace patterns with ANN-related tools offers novel opportunities to conceptualize the "potential" for an individual to cause harm within critical infrastructure. Where Merriam-Webster's Online Dictionary defines potential as "something that can develop or become actual," there is a need to align with—and build upon—traditional ITDM characteristics. Regardless of how traditional preventive and protective measures for ITDM are interpreted (consider, [12] and [13], for example), much of the relevant literature agrees that insiders often reveal evidence of their malicious acts *prior* to executing their attack (e.g., The Fort Hood terrorist attack) [15]. Working toward a flexible conceptual approach and agnostic view on specific insider pathways, incorporating facility-level signals related to operational patterns and workplace rhythms forms the foundation for monitoring the potential actualization of insider acts.

In response, the concept of *risk significance*—borrowed from a long, successful history in the nuclear safety domain—provides a strong structure for crystallizing the "potential" for an insider act. Explained in depth in [15], this concept relates to characterizing estimated frequencies of adverse events and degrees of consequence against an anticipated baseline, where exceeding a policy-determined limit is definitionally risk significant. If the risk of an insider act is calibrated against a threshold-defining acceptability, then insider potential can be described as a time-variant continuous variable describing operational patterns that move toward (or away) from this threshold. From this perspective, critical thresholds for risk potential will vary by facility and the ability (e.g., access, authority, and knowledge) for an individual to execute a malicious act. Thus, in general, individuals that currently possess the ability to execute a malicious act are considered "highly" risk significant, where individuals without that ability are typically "highly" risk *in*significant. Yet, the proposed ITDM framing also incorporates the interdependencies between organizational dynamics and workplace rhythms—allowing for data gathered by ANNs over time from facility sensors to categorize measures of operational patterns and workplace rhythms into levels of risk significance at the individual-level both before and during an actual insider attack/theft.

For example, consider the ability of current critical infrastructure policies and systems to detect genuine insider threats before or after any malicious action is executed. Those efforts aimed to detect insider potential *before* a malicious act is completed can be called *Type I* detection, while efforts aimed at detection *after* a malicious act can be called *Type II* detection. An extremely "high" risk significant insider is an individual capable of avoiding both types of detection to successful execute a malicious act—indicating that baselines or thresholds for *individual* risk significance are related to *organizational* efforts in Type I and II detection. Here, incorporating ANNs can help monitor collective behavior via workplace rhythms, compare the data to expected individual-level behavior, highlight deviations from expectation behaviors, and identify anomalous behaviors near (or past) established thresholds built on Type I and Type II detection. In short, the larger the deviation or anomaly, the higher the insider risk significance or insider potential.

Consider the following vignette to clarify this overall logic:

Jane is an advanced undergraduate student who regularly works within her university's nuclear research reactor. One day, she is invited by her academic advisor to observe a graduate research project. This graduate research is conducted in a non-controlled (but sensitive) reactor area in which Jane has not worked before. Over a period of several weeks, Jane visits this other research project on several occasions (during her normal working hours), but neither Jane nor her professor alert facility security of this expected change.

While sensors do not register a new time pattern for Jane entering the research reactor, her movements in new areas of the facility are registered by various physical sensors as what appears to be a deviation from her expected workplace rhythms. In this case, Jane's risk significance has increased as a result of the combined changes in her behavior (accessing new facility areas) and a failure of policy (security was not alerted by her temporary change in role). Given that both of these elements—which align to Type I and Type II detection—are measurable, risk significance can be expressed in terms of anomalies from expected behaviors. With ANNs able to ingest data related to individual workplace rhythms, risk significance (and insider potential) could be monitors as deviations from expected behavior in (near) real-time.

While traditional ITDM approaches for critical infrastructure facilities have improved in recent years, many are still driven by peer-to-peer reporting and individual behavioral observation mechanisms. Yet, shifting focus from such individually focused interpretations of insider opportunity to facility-focused insider potential helps address the need for a new insider threat framework that utilizes advances in data analysis to better characterize deviations from expected operational patterns and workplace rhythms. If insider opportunity is often considered a function of individual access, authority, and knowledge in traditional approaches, then a new approach would contextualize access, authority, and knowledge in terms of collective behaviors emerging from how individual interact with organizational ITDM efforts. Leveraging ANNs to assist with data collection and evaluation, such a collective behavior-based ITDM approach could help overcome issues related inaccurately conflating human error with malicious acts, adequately attributing motivational triggers to insider acts, and communicating anomalous behavior within a facility to initiate a proper response. Estimating expected workplace behavior and measuring deviations from expected behaviors in terms of Type I or Type II detection provide an enhanced ability to detect and mitigate "the *potential* for an insider to… harm that organization" (emphasis added) [5].

3 METHODS & DATA COLLECTION

To further explore this new approach to ITDM for critical infrastructure, empirical data was needed to capture operational patterns and workplace rhythms that emerge as individuals settle into routines for executing job-related tasks. Such patterns and rhythms can be described via data signals *already collected* at many critical infrastructures, including (nut not limited to) access control, intrusions sensor, camera video, area or personal safety monitoring, and material control data. The extent to which empirical bounds to these patterns and rhythms exist within these data sources supports the ability for both expected and anomalous operational behaviors to be identified—resulting in framework for measuring risk significance and insider potential. Organization B provided a representative critical infrastructure facility within which to investigate this data-driven, collective behaviors-based ITDM approach. NETL—which hosts the newest TRIGA Mark II nuclear research reactor in a U.S. university—is an innovative facility with unique capabilities and a multifaceted mission that includes educating next-generation leaders in nuclear science and engineering; performing cutting-edge research across several primary thrusts (including, but not limited to, nuclear forensics, trace element analysis, neutron depth profiling, and radiography imaging); and producing radioisotopes for research, nuclear medicine, and industrial processes.

Consistent with an exploratory approach, NETL hosts a range of personnel—including operational and administrative staff; faculty; post-doctoral and staff researchers; graduate and undergraduate students; contractors, and visitors—whose patterns and rhythms may be described by already collected access control and intrusion detection data signals. Where access control data supports understanding when and how frequently authorized personnel enter a given area (as well as attempts at unauthorized access), intrusion detection data supports understanding the implications of registered movement in protected areas under different NETL operational states. Collected as independent data streams, simple statistical tests could be compared against externally generated thresholds or expectations. Yet, incorporating ANNs to evaluate them as interdependent data streams captures more complex and nuanced scenarios leading to a deeper understanding of expected patterns and rhythms. Table 1 summarizes the sensor and data type collected by the ANN, as well as corresponding activities captured by each sensor at a facility such as NETL. All collected data was anonymized following best research practices [16], including using genericized average data values, invoking facility personnel categories, and performing analysis at higher levels of data aggregation. (Note: If such an ANN were operationally deployed, then individual identities would be another data stream incorporated into the learning process—and ITDM mitigation strategies.)

Table 1: Description and categorization of data related to a representative set of expected organizational activities at NETL, recreated from [11].

ITDM Category	Sensor Type	Data Type	Representation Organizational Activity
Access Control	• Badge reader ▪ NETL entry ▪ Security control panel ▪ Limited area ▪ Reactor control room	• Badge readers: ▪ # authorized attempts ▪ # unauthorized attempts (false negative + false positives) ▪ Time of access attempts	• Personnel arrival to facility • Researchers approaching the reactor • Reactor operator arriving for shift
Intrusion Detection	• Balanced magnetic switch ▪ Limited area ▪ Security control panel ▪ Reactor control room	• Balanced magnetic switches: ▪ # times switch opened ▪ Time at which switch opens	• Researchers approaching the reactor • Maintenance of security control panel • Reactor operator arriving for shift
	• Area motion sensor ▪ Reactor bay ▪ Fuel storage surveillance	• Area motion sensors: ▪ # times change in physical phenomena registered ▪ Time at which change in physical phenomena registered	• Custodial services around the reactor • Transfer of fresh/used fuel into/out of NETL

Several different commercially available ANN (or ANN-adjacent) tools were used to capture and evaluate the various data sets related to operational patterns and workplace rhythms at NETL. Each of the tools collected information from facility sensors for a period of time in order to generate a training dataset, evaluated that training data until an internal performance metric (e.g., validation error) was achieved, and then optimized their respective algorithms (e.g., via stochastic gradient descent). Tool 1 integrated into NETL's existing security posture and implemented role-based, risk-adaptive access controls to ensure that all authorized personnel only visit the appropriate areas. Similarly, Tool 2 utilized existing security mechanisms and the proprietary algorithm to identify changes to operational behavior by facility personnel. (Note: The focus of this project is to explore the new ITDM approach, so no formal comparison between the Tools was conducted.) Each of these tools—according to their own proprietary algorithmic structure— collected and evaluated data from sensors already deployed to learn the flow of people and processes at NETL and monitored for insider potential by managing risk and flagging and reporting anomalous events.

In total, four data sets were collected and evaluated across the different ANN (and ANN-adjacent) tools. The data sets incorporate several interested elements. First, different data sets align to different tools. Second, data collected for Tool I occurred over three disparate time periods, which coincidentally aligned with the different set of operational patterns and workplace rhythms observed (and experienced) “before” and “after” COVID-19 precautions took place in 2020. Third, the different time periods of data collection for the different tools affords the opportunity to evaluate the ability for operational patterns and workplace rhythms to evolve naturally *without* triggering deviation or anomaly of concern. Lastly, data points were loosely organized into two categories to observe trends in patterns and rhythms at NETL. The first category consisted of single-access-points (SAP). Organized by access point, date and time of allowed access, and identity used for access, SAPs used sensor observations to produce ANN-reported patterns and time bounds on expected general access for an average (and specific) individual. The second category consisted of time-sequenced, multiple-access-points (TSMAP). Similarly organized to SAPs, TSMAPs used sensor observations to produce ANN-reported patterns to identify time bounds on expected completion of a particular access sequence for particular individuals. By extension, deviations in SAPs manifest as attempted accesses outside of these empirically defined time bounds, where deviations in TSMAPSs manifest as either attempted accesses outside of empirically defined time boundaries or attempted accesses in a different order than in empirically defined baseline patterns. Table 2 summarizes the data collected at NETL.

Table 2: Summary of artificial neural network data collected from NETL, updated from [11].

Data Characteristic	Data Set I	Data Set II	Data Set III	Data Set IV
ANN Solution	Tool 1	Tool 1	Tool 1	Tool 2
Date range	10/12/2019 to 03/14/2020	03/15/2020 to 09/25/2020	09/26/2020 to 03/31/2022	03/15/2023 to 09/15/2023
Access control data points	13,653	18,986	74,922	27,653
Intrusion detection data points	694	923	4211	1102
Categories for organizing data points[a]	SAP TSMAP	SAP TSMAP	SAP TSMAP	SAP TSMAP

[a] SAP = single-access-point operational patterns; TSMAP = time-sequenced, multiple access point operational patterns

4 DEMONSTRATING A NEW APPROACH TO ITDM

Results (shown below) provide examples of how ANN tools capture and illustrate how access control and intrusion detection signals collected at NETL describe operational patterns and workplace rhythms. While this ANN-approach is capable of identifying patterns and rhythms for individuals, this research project focused on personnel categories to emphasize the explanatory power of collective behaviors and avoid potential privacy issues. Summarized around personnel categories, the data shows a (somewhat surprising) level of regularity in certain patterns and rhythms. For example, consider the SAP-based frequency distribution of the first allowed access to NETL versus the time of day across data sets in Figure 1. This visualization helps identify and define category-specific profiles of expected first entry to the facility (as well as other access points throughout the facility).

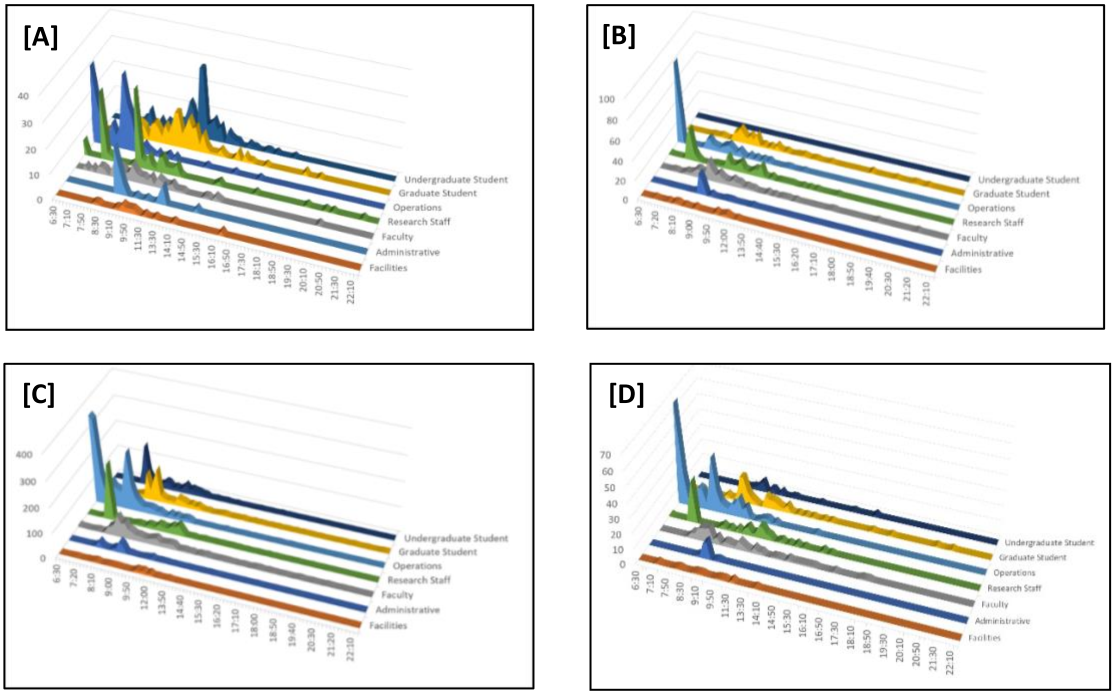

Figure 1: SAP frequency distribution showing time of first entrance to NETL separated by personnel group for [A] Data Set I (Tool 1 outputs before COVID-19 lockdowns), [B] Data Set II (Tool 1 outputs immediately after initial COVID-19 lockdowns), [C] Data Set III (Tool 1 outputs from late 2020 to 2022), and [D] Data Set IV (Tool 2 outputs), updated from [17].

In some cases, these patterns are very tightly bounded in time (for example for the administrative and operational personnel) and in other cases these patterns have wide distributions (for example the faculty, undergraduate students, and graduate students). Comparing across the results from the four data sets suggests that (1) collected data signals *can* reflect patterns and rhythms in behaviors, (2) common patterns and rhythms can form profiles associated with particular personnel categories, and (3) such personnel category profiles can be used as a baseline of expected behaviors—each of which supports a data-driven, collective behavior-based framework to detect deviations that may reflect insider potential insider.

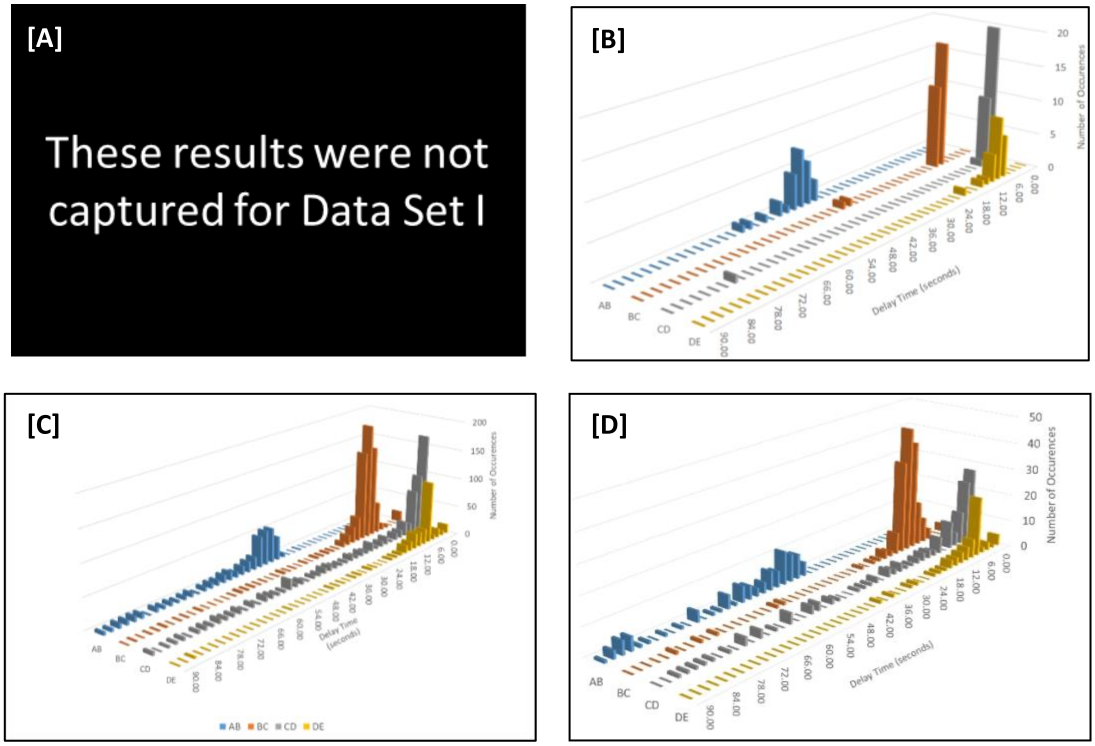

Figure 2: TSMAP frequency distribution showing time-series plot of time delays between authenticated access along a series of entry points (A, B, C, D, E) that compose the expected behavior of the first person to enter the facility and head toward the research reactor control room, where [A]-[D] correspond to Data Sets I-IV, respectively.

Similarly, Figure 2 illustrates the TSMAP results related to the time taken between accessing entry points along a given path through NETL. Across the four data sets, the ANNs registered observations to determine expected patterns and rhythms emerging from moving through the access points A, B, C, D, and E through NETL—resulting in an emergent, dynamic patterns of expected behavior. The results shown in Figure 2 provide a higher fidelity and more nuanced depiction of expected patterns and rhythms that share—and enhance—the benefits of using such data-driven profiles to establish baselines for expected behaviors. More specifically, the frequency distributions across the data sets would allow such statements as “this individual is expected to take 42-66 seconds to move from access point A to access point B” (Data Set II) or “this individual is expected to take 6-30 seconds to move from access point C to access point C” (Data Set III). The ability to empirically derive these types of statements add validity and structure to using deviations from expected patterns to describe insider potential as the underlying logic of this new ITDM approach.

In addition, several scenario-based experiments were conducted to further explore this ANN approach to ITDM, which included different potential insider pathways for executing malicious acts on three different target locations within NETL—the closet housing the security system’s control panel, the reactor bay, and nuclear fuel storage area. For clarity, the scenario names were matched to the potential target location and insider pathways refer to increasingly sophisticated strategies taken to execute an action. The first—designated insider pathway A—incorporates attempting direct access by an unauthorized individual, including single and multiple attempts during both normal and off hours. The second—designated insider pathway B—corresponds to attempting direct access by an unauthorized individual during off- hours using credentials from an authorized individual, but who used their own credential to enter the NETL building. Lastly, designated insider pathway C depicts attempting direct access by an unauthorized individual during off- hours using credentials from an authorized individual and using that same credential to enter the NETL building.

Table 3: Summary description of experimental results for insider threat-related scenario analysis, updated from [11].

Scenario Name (#)	Test Description (Scenario # & Pathway Name)	Data Set I Results*	Data Set II Results	Data Set III Results	Data Set IV Results
Security Closet Access (1)	Unauthorized Access Attempt (1A)	Detected & Denied in ***ALL*** Cases [SAP]	Detected & Denied in ***ALL*** Cases [SAP]	Detected & Denied in ***ALL*** Cases [SAP]	Detected & Denied in ***ALL*** Cases [SAP]
	Authorized Access Credentials Used by Unauthorized Individual Who Entered Building Using Their Own Credentials (1B)	Detected & Denied in ***MOST*** Cases [SAP; TSMAP]	Detected & Denied in ***MOST*** Cases [SAP; TSMAP]	Detected & Denied in ***MOST*** Cases [SAP; TSMAP]	Detected & Denied in ***NO*** Cases [SAP; TSMAP]
	Authorized Access Credentials Used by Unauthorized Individual Who Entered Building Using Authorized Individual's Credentials (1C)	Detected & Denies in ***NO*** Cases [TSMAP]	Detected & Denies in ***NO*** Cases [TSMAP]	Detected & Denies in ***MOST*** Cases [TSMAP]	Detected & Denied in ***MOST*** Cases [SAP; TSMAP]
Reactor Bay Access (2)	Unauthorized Access to Reactor Bay (2A)	Detected & Denied in ***ALL*** Cases [TSMAP]	Detected & Denied in ***ALL*** Cases [TSMAP]	Detected & Denied in ***ALL*** Cases [TSMAP]	Detected & Denied in ***ALL*** Cases [TSMAP]
	Early Detection by Motion Sensor (2B)	Not Tested	Detected in ***MOST*** Cases	Detected in ***MOST*** Cases	Detected & Denied in ***NO*** Cases [SAP; TSMAP]
Fuel Storage Surveillance (3)	Insider Surveillance (3A)	Difficult to Detect Without Additional Sensing Input [TSMAP]	Difficult to Detect Without Additional Sensing Input [TSMAP]	Difficult to Detect Without Additional Sensing Input [TSMAP]	Detected & Denied in ***NO*** Cases [SAP; TSMAP]
	Insider Alarm Testing (3B)	Not Tested	Difficult to Detect Without Additional Sensing Input [TSMAP]	Difficult to Detect Without Additional Sensing Input [TSMAP]	Detected & Denied in ***NO*** Cases [SAP; TSMAP]

*SAP = single-access-point operational patterns; TSMAP = time-sequenced, multiple access point operational patterns

Using a combination of SAP and TSMAP data, the ANN tools were evaluated on their ability to register unauthorized access attempts as anomalous behaviors. (NOTE: For more detailed description of the analysis of data sets I and II, please see [11].) Consider, for example, how Scenario (1) evaluated data collected from access control readers at the facility entrance and near security-related control systems. The SAP results yielded successful ANN detection and denial of access across all data sets for test (1A)—which matches intuition and benchmarks this ITDM approach to well established capabilities (and performance) of traditional access control systems. The other two hypothesized insider pathways for

Scenario 1 were similarly evaluated, where Tool 1 improved in its ability to detect and deny across data sets I and III, while Tool 2 was able to detect and deny unauthorized access in most cases for test (1C) but in no cases for test (1B).

For Scenario (2), data was collected from multiple access control readers leading to the NETL reactor bay, as well as motion sensors within the reactor bay itself. In this scenario, the ANN tools acted as ITDM systems looking for off-normal activity that would include both attempts at unauthorized access (similar to Scenario 1) and early detection of the insider moving toward the reactor bay. And, similar to the results summarized above, TSMAP results yielded successful ANN detection and denial of access across all data sets for test (2A). While Tool 2 did not successfully detect and deny unauthorized access in test (2B), the increased success experienced by Tool 1 between data sets I and III suggest additional learning can increase Tool 2's future success. Lastly, Scenario 3 collected from motion detection sensors, reactor bay access controls, intrusion detection sensors, area radiation sensors, and alarm panel sensors to evaluate potential surveillance activities. Here, test (3A) aimed to explore the ability of an ANN to detect evidence of insider surveillance activities within the reactor bay and test (3B) aimed to assess an ANN's ability to detect insider testing of access alarms for nuclear fuel storage. Both of the ANN tools struggled significantly with these tests, despite the additional data gathered between data sets II and III. These results do suggest a need for additional data sources to more appropriately characterize baselines of operational patterns and workplace rhythms against which to identify possible deviations that could be associated with potential insider actions. Table 3 shows the experimental results for each scenario across the four data sets.

Results from the experimental efforts to determine empirically validated baselines for expected behaviors and explore different insider scenarios support the risk significance approach to managing ITDM efforts. For example, the extent to which ANN tools can reflect actual behaviors describing how individuals interact with organizational ITDM policies and systems indicates the opportunity to establish baselines that can inform thresholds separating risk significant from risk *in*significant insider potential. The SAP-based (Figure 1) or TSMAP (Figure 2) profiles provide the scaffold for characterizing functionally unacceptable behaviors (e.g., risk significant insider potential) as a quantified deviation from expected behaviors. In addition to the benefits described above, the scenario experimentation also provides the ability to conduct sensitivity analysis on the ANN tools. This capability can help derive tailored risk significance thresholds for different facility personnel categories to support a more comprehensive and holistic insider potential framework.

More specifically, consider the risk significance framework offered in Figure 3. If risk significance relates to the success of organizational policies and systems to detect an insider before (Type I) and after (Type II) a malicious act, then thresholds exist that demarcate when observed behaviors transition into being insider risk significant. In Figure 3, the red and blue dashed lines represent such thresholds. This conceptualization offers several beneficial characteristics. First, the thresholds can be (at least partially) derived from profiles of expected behaviors collected by ANNs from operational patterns and workplace rhythms. Second, different thresholds can be illustrated on the same risk significance framework, improving efficiency in ITDM program management. Third, different facility personnel categories will align differently across the thresholds—allowing more tailored policies and systems to be designed and executed. Lastly, mapping either individuals or facility personnel categories on this framework also supports *anticipatory* ITDM—where indications of increasing insider potential can initiate an organizational response prior to an executed malicious act.

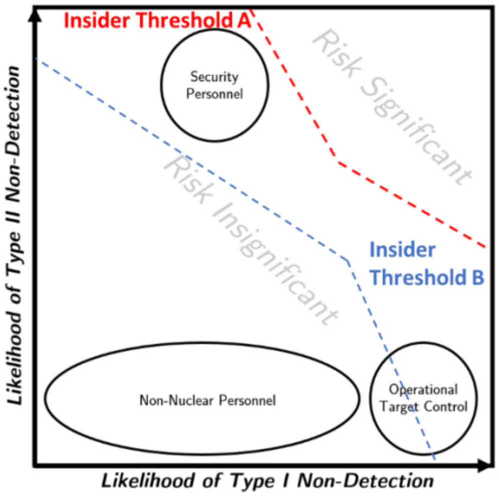

Figure 3: Notional risk significance framework (modified from [15]) illustrating insider potential in terms of the relative success of Type I (*before* a malicious act) and Type II (*after* a malicious act) detection efforts established by an organization, where two representation insider thresholds (which can be informed by operational patterns and workplace rhythms) are represented by red and blue dashed lines.

5 CONCLUSIONS, INSIGHTS & IMPLICATIONS

Conclusions on the efficacy of the proposed data-driven, collective behavior-based approach to ITDM are gleaned from an exploration that encompassed two ANN tools; four data sets; describing operational patterns and workplace rhythms with various signals; scenario analysis for increasingly sophisticated insider pathways; and mapping insider threat thresholds on a risk significance framework. Consistency in aggregating access control and intrusion detection signals into SAP and TSMAP outputs across data sets indicate that ANNs can effectively identify patterns of expected workplace behaviors and translate them into baseline profiles against which to investigate if anomalies are indicative of increasing insider potential. Such behavioral profiles afford opportunities to quantify deviation from expected behaviors that would allow related ITDM approaches to objectively investigate anomalies and better distinguish between incidental deviations from those portending increased insider potential. Trends in scenario analysis across data sets demonstrate the flexibility of this data-driven, collective behavior-based ITDM framework that employed different ANN tools with increasing numbers of data streams to effectively detect and deny increasingly sophisticated insider pathways for causing malicious hard at several target areas. Though the most unrefined element of this ITDM approach, the risk significance framework offers clear, transparent, and informative common mental map of insider potential that helps overcome biases often related to ITDM program that heavily emphasize mitigating individual psychological stressors/indicators.

Synthesizing these experimental conclusions also generated several insights related to enhancing ITDM in critical infrastructure. For example, consider how the ability for ANNs to "learn" operational patterns and workplace rhythms translates into a capability to dynamically match regularly experienced (and observed) natural evolution in facility operations. This capability ultimately reduces "false positives" by better distinguishing deviations caused by natural evolution (e.g., the difference between data sets I and II in Figure 1) and those potentially related to increased insider potential (e.g., a registered a deviation from the expected norm in any of the data sets in Figure 1). Another insight drawn from the results demonstrate how ANN (and ANN-adjacent) tools improve their ability to identify increasingly sophisticated, complex, or nuanced insider pathways, as highlighted by the improved performance of Tool 1 in sensitivity analysis across data sets I, II, and III. Similarly, consider how incorporate risk significance helps coordinate the new data, technical, and conceptual characteristics proposed in this ITDM approach—particularly in offering an enhanced, streamlined investigation and communication capacity. Lastly, a dynamic interpretation of risk significance (as introduced in [15]) further suggests an ability to anticipate which future deviations in patterns and rhythms are most likely indicate malicious intent to improve ITDM management.

Though there is still a fair amount of analytical (e.g., validation and verification) ground to cover, the results to date from the proposed data-driven, collective behavior-based ITDM approach offers a few interesting implications for critical infrastructure. First, this approach provides an operationally "lean" ITDM program that only requires an ANN (or ANN-adjacent) tool that will gather information from signals *already collected* onsite and coordinate them into meaningful ITDM-relevant outputs to aid decision making. Second, adding data sources—including both generic data sources (e.g., camera data) and sources specific to a given industry (e.g., personal radiation dosimeters at nuclear facilities)—would continue to improve the analytic power of ITDM approach to better detect and mitigate increasingly sophisticated, complex, and nuanced insider acts. Lastly, this shift toward an "insider potential" framing offers an empirically supported ITDM approach that quantitatively describes operational patterns and workplace rhythms in terms of risk significance—which represents cutting-edge advances that support both domestic (e.g., U.S. NITTF programs) and international (e.g., International Atomic Energy Agency's INFCIRC/908) efforts to enhance ITDM.

ACKNOWLEDGMENTS

The authors would like to thank the National Nuclear Security Administration's Office of International Nuclear Security (NNSA/INS) for funding this study. We also acknowledge the feedback and insights gained from discussions with various colleagues across the NNSA laboratory complex and members of the Institute for Nuclear Materials Management. Lastly, we thank each of our reviewers for insightful comments. Sandia National Laboratories is a multimission laboratory managed and operated by National Technology & Engineering Solutions of Sandia, LLC, a wholly owned subsidiary of Honeywell International Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525.

REFERENCES

[1] National Insider Threat Task Force. 2016. Protect Your Organization from the Inside Out: Government Best Practices. NITTF Report

[2] Adam D. Williams., Shannon N. Abbott, and Adriane C. Littlefield. 2019. Insider Threat. In Encyclopedia of Security and Emergency Management, eds. L. Shapiro and M.H. Maras, Springer, Cham

[3] Federal Register Vol. 76, No. 198. 2011. Presidential Documents. Retrieved from: https://www.dni.gov/files/NCSC/documents/nittf/EO_13587.pdf

[4] National Insider Threat Task Force. 2017. Insider Threat Guide: A Compendium of Best Practices to Accompany the National Insider Threat Minimum Standards. Washington, DC

[5] Cyber and Infrastructure Security Agency. 2020. Insider Threat Mitigation Guide. Washington, DC. Retrieved from: https://www.nationalinsiderthreatsig.org/itrmresources/CISA%202020%20Insider%20Threat%20Mitigation%20Guide.pdf

[6] International Atomic Energy Agency. 2008. Preventive and Protective Measures Against Insider Threats. IAEA Nuclear Security Series No. 8: Implementing Guide

[7] Carolynn P Scherer and Christy E. Ruggiero. 2019. Overview of Tools for Insider Threat: Analysis and Mitigation. Retrieved from: https://permalink.lanl.gov/object/tr?what=info:lanl-repo/lareport/LA-UR-19-22069

[8] World Institute for Nuclear Security. 2018. Countering Violent Extremism and Insider Threats in the Nuclear Sector

[9] Richard M. Cyert and James G. March. 1963. A Behavioral Theory of the Firm (2nd. Ed.). Prentice-Hall, Malden, MA.

[10] Anthony Giddens. 1984. The Constitution of Society: Outline of the Theory of Structuration. University of California Press, Berkley, CA.

[11] Adam D. Williams, Shannon N. Abbott, Nathan Shoman, and William S. Charlton. 2021. Results From Invoking Artificial Neural Networks to Measure Insider Threat Detection & Mitigation. Digital Threat: Research and Practice 3, 1, Article 3.

[12] Mark F. Lenzenweger and Eric D. Shaw. 2022. The Critical Pathway to Insider Risk Model: Brief Overview and Future Directions. Counter-Insider Threat Research and Practice 1, 1.

[13] Jordan Richard Schoenherr, Kristoffer Lilja-Lolax, and David Gioe (2022). Multiple Approach Paths to Insider Threat (MAP-IT): Intentional,

Ambivalent and Unintentional Insider Threats. Counter-Insider Threat Research and Practice 1, 1.

[14] Amy B. Zegart. 2017. The Fort Hood Terrorist Attack: An Organizational Postmortem of Army and FBI Deficiencies. In Insider Threats, ed. Matthew Bunn and Scott Sagan, pg. 42-73. Cornell University Press

[15] Christopher A. Faucett. 2022. Development of a Conceptual Multi-Insider Risk Model for Nuclear Facilities. PhD Thesis, Nuclear Engineering Department, Texas A&M University

[16] Robert K. Yin. 2016. Qualitative Research from Start to Finish, 2nd Edition. Guilford Press, New York, NY.

[17] Colton Heffington, Adam D. Williams, Shannon Abbott, Christopher Faucett, Sondra Spence, William Charlton, Katherine Holt, and Melinda Lane. 2023. Operationalizing Insider Threat Potential and Risk-Significant Insiders to Enhance Insider Threat Detection and Mitigation. Proceedings of the Annual Meeting of the Institute of Nuclear Materials Management.