A System Dynamics Model for Investigating Early Detection of Insider Threat Risk

Andrew P. Moore, apm@cert.org, 412-268-5465
David A. Mundie, dam@cert.org, 412-268-3552
Matthew L. Collins,[1] mlcollins@cert.org, 412-268-9152

CERT®[2] Program Software Engineering Institute
Carnegie Mellon University
4555 Fifth Avenue
Pittsburgh, PA 15213

Abstract

In many organizations, the responsibility for managing the insider threat falls almost exclusively with the information technology staff. But many of the early indications of problems occur at a behavioral, nontechnical level. This paper describes a system dynamics model for investigating how monitoring the behavioral indicators of insider threat risk can reduce the overall risk of a cybersecurity breach within an organization by promoting early detection. We show how the model could be used for a given set of input data, derived from our insider threat case database, and discuss future work to identify more robust inputs through interaction with partnering organizations.

1 Introduction

The 2011 CyberSecurity Watch survey revealed that 27%[3] of cybersecurity attacks against organizations were caused by disgruntled, greedy, or subversive *insiders*:[4] employees or contractors with access to the victim organization's network systems or data. Of the 607 survey respondents who knew about the relative financial impact of insider and outsider attacks, 46%[5] viewed insider threat attacks as more costly than attacks from the outside, usually in terms of financial loss, damage to reputation, critical system disruption, and loss of confidential or proprietary information (CSO Magazine, U.S. Secret Service, SEI CERT, Deloitte, 2011). In both the U.S. Department of Defense (DoD) and industry, insiders' authorized physical and logical access to organizational systems and their intimate knowledge of the organizations make combating insider threat attacks difficult. Unfortunately, current countermeasures to insider threat are largely reactive, leaving information systems that contain sensitive information susceptible to the procedural and *technical* vulnerabilities commonly exploited by insiders.

---

[1] At the time this paper was written, Matthew Collins was a graduate assistant at CERT and a graduate student at the H. John Heinz III College at Carnegie Mellon University. He is now member of the technical staff at the CERT Program.

[2] CERT and CERT Coordination Center are registered marks owned by Carnegie Mellon University.

[3] This percentage is of those incidents in which the respondent knew whether an insider or an outsider was responsible for the attack. Respondents did not know in 21% of all incidents.

[4] Terms that are defined in Appendix B: Glossary appear in italics on first use in the text.

[5] This percentage is of those respondents who knew whether insider or outsider attacks were more costly to their organization. Respondents did not know in 29% of all incidents.

Insider threat detection is an essential element of any insider threat program. This paper addresses the difficulty of early detection of malicious insider threat risk.[6] Most organizations only detect that they are at risk of insider compromise after they have been attacked. Earlier detection would allow organizations to prevent or limit the impact of an attack. Unfortunately, the most often observed indicators of risk occur very late in the lifecycle of the incident.

Over time, an insider may engage in certain behaviors that indicate, based on defined decision rules, increased risk of malicious attack. Figure 1 presents an example of an employee's potentially malicious behavior, its observation, and the timing of the organization's response. The lower line represents the rising indication of risk observed by the organization. Time 0 indicates the point at which insiders are hired; they start at a positive risk value because of personal predispositions they bring to the job. The organization takes action when the observed risk rises above the alert threshold.

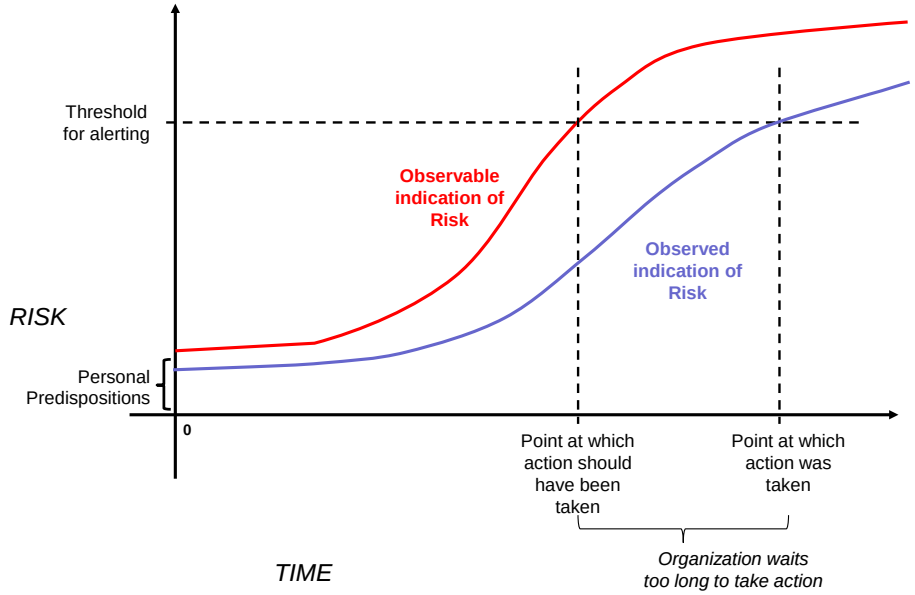

Figure 1: Example Observation of Insider's Risk to an Organization

The upper line represents the *observable* indication of the risk posed by the insider. This is the sum total of all of the employee's actions that indicate risk. The more the organization knows about the insider's actions, the sooner the insider would cross the organization's alert threshold.

The primary point to consider is the difference between the observed indication of risk and the observable indication of risk. The point at which the organization took action (indicated by the time at which the lower line crosses the alert threshold) is substantially later than the point at which action could have been taken (indicated by the time at which the upper line crosses the alert threshold). This delay in taking action could be the difference between responding to an insider attack and preventing it. At the least, earlier detection will allow improved mitigation of

[6] In this paper, insiders include current or former employees, contractors, and other business partners—anyone with authorized access to an organization's systems beyond that provided to the general public. Malicious insider threat is the potential harm from insiders intentionally using or exceeding their authorized access in a way that damages the organization.

the attack, possibly limiting the damage to the organization. Perfect knowledge is unattainable; even if gathering all indicators of risk is possible, the cost of gathering and analyzing all of the indicators might outweigh the benefits.

In an operational environment, it is difficult to control for all the factors that might influence insider threat detection measures. In addition, we do not generally know the distribution or frequency of actual insider attacks due to deficiencies in detection capabilities and in the reporting or attribution of malicious insider activities. To address these concerns, our work will use system dynamics modeling and simulation to identify and control as many factors as possible in a closed test environment.[7] System dynamics helps analysts model and analyze critical behavior as it evolves over time within complex socio-technical domains. (Sterman, 2000) (Meadows, 2008) A powerful tenet of this method is that the dynamic complexity of critical behavior can be captured by the underlying feedback structure of that behavior. The boundaries of a system dynamics model are drawn to encompass all the enterprise elements necessary to generate and understand problematic behavior. An overview of the system dynamics modeling notation is provided in Appendix A.

This paper provides the foundation for work to determine the potential for earlier detection of heightened risk due to insider threat. The ultimate objective will be to demonstrate that considering a broad range of insider threat indicators—both behavioral and technical—in risk analysis will significantly help to either prevent insider compromise or detect it as early as possible. We believe that if an organizations' decision makers recognize, record, and quickly relay to insider threat analysts a broader range of behavioral and technical indicators (identified in the CERT database), the organization can drastically hasten its detection of a significant insider threat risk, perhaps by as much as one-third to one-half of current detection times.

2 Related Work

Research on insider threat can be broadly characterized as one or more of the following:

- case based—investigating and analyzing insider threat incidents as the basis for understanding the problem and effective solutions
- experimental—conducting scientific experiments to test hypotheses about insider activities
- detection—detecting malicious insider acts early in the lifecycle of the crime so forestalling action can be taken
- prevention—preventing the malicious acts in the first place, independently of detection

Related research in the area varies by how much technical and non-technical aspects of the crime are considered. Work focusing primarily on non-technical aspects includes the work at the University of Nebraska was strictly case based and detection focused (Bulling, Scalora, Borum, Panuzio, & Donica, 2008). Sandia's work (Duran, Conrad, Conrad, Duggan, & Held, 2009) is limited to system dynamics simulations of process-oriented countermeasures identified by studying the employee lifecycle. General deterrence theory is a framework limited to preventing crime generally (Straub, 1986).

[7] We use the VenSim environment by Ventana Systems, Inc: http://www.vensim.com.

The seminal Advanced Research and Development Activity (ARDA) report (Maybury, et al., 2005) documents the most intensive technical examination of insider threat, but it was mostly detection based. The emerging Defense Advanced Research Projects Agency (DARPA) Cyber-Insider Threat (CINDER) work (DARPA Strategic Technology Office, 2010) is also more technically oriented but appears to include both preventive and detective approaches.

Much of the research community recognizes that both organizational and technical issues need to be addressed for insider threat defense. The emerging DARPA Anomaly Detection at Multiple Scales (ADAMS) work (DARPA Information Innovation Office, 2010) focuses primarily on insider threat detection. Case-based approaches include work at the DoD Personnel Security Research Center (Shaw & Fischer, 2005) and at RAND (Predd, Pfleeger, Hunker, & Bulford, 2008), and work on insider attack surfaces (Blackwell, 2009).

A few efforts involve experiments: notably the work at MITRE (Caputo, Stephens, & Maloof, 2009) and Pacific Northwest National Laboratory (PNNL) (Greitzer, et al., 2009); and the emerging Defense Advanced Research Projects Agency (DARPA) Anomaly Detection at Multiple Scales (ADAMS) research (DARPA Information Innovation Office, 2010).

A few efforts are strictly prevention focused, including the insider attack surface research (Blackwell, 2009) and work on the theory of situational crime prevention (Willison & Siponen, 2009). Several efforts involve both detection and prevention approaches (Predd, Pfleeger, Hunker, & Bulford, 2008) (Shaw & Fischer, 2005) (Caputo, Stephens, & Maloof, 2009), but they are either strictly case based or experimental.

The literature stops just short of expressing an explicit temporal relationship between concerning non-technical and technical observables in insider IT sabotage crimes. However, several sources strongly suggest that non-technical observables precede technical observables. For example, Shaw describes the insider crimes along a Critical Pathway that shows personal and professional stressors followed by maladaptive behavioral reactions (often conflict at work), followed by management intervention that fails to divert (and possibly even escalates) the attack. (Shaw & Fischer, 2005) Elements of this pathway are also evident in the psychopathology literature, in particular the Diathesis-Stressor Model and Cascade Modeling.[8] (Monroe & Simons, 1991) (Ingram & Price, 2001) Given that the attacks are typically the most technical aspects of the crime, it is natural to assume that the technical observables come later along the pathway than the non-technical observables.

The U.S. Secret Service joint work with the CERT Program in a study of insider computer system sabotage showed that "a negative work-related event triggered most insiders' actions" and "most of the insiders had acted out in a concerning manner in the workplace." (Keeney, et al., 2005) From a technical perspective, that study found that "the majority of insiders compromised computer accounts, created unauthorized backdoor accounts, or shared accounts in their attacks." Again, it seems logical to infer from these findings, that the non-technical aspects of the timeline occurred before the technical. However, this is just an inference not a firm conclusion.

[8] The August and November issues of the 2010 volume of the Cambridge journal Development & Psychopathology are devoted to cascade modeling.

3 A Qualitative Model of Insider Threat Detection

System dynamics and the related area of systems thinking encourage the inclusion of soft factors in the model, such as policy, procedural, administrative, or cultural factors. The exclusion of soft factors in other modeling techniques essentially treats their influence as negligible, which is often an inappropriate assumption. This holistic modeling perspective helps identify mitigations to problematic behaviors that are often overlooked by other approaches. This section provides a qualitative model of important aspects of the insider threat detection problem.

In general, at an organization level, employees are monitored but incident investigators only get two types of information as a result: (1) technical information reported by automated sensors and (2) behavioral information reported through "human sensors" throughout the organization. At a local (unit) level, immediate supervisors are arguably the most important human sensor within an organization.

The causal loop diagram in Figure 2 illustrates how insider threat detection is supported through supervisor reporting. Starting at the bottom of the diagram and moving counterclockwise, we see that the variables of insider opportunity and insider incentive (shown in maroon, italics) spur the insider activities related to the crime. Moving along the right side of the diagram, this spurring increases supervisors' opportunity to detect the insider threat, as well as their knowledge of indicators, provided they are appropriately monitoring for those indicators. Along the top of the diagram, if supervisors are willing to report suspicious insider behaviors, insider threat indicators will be relayed to analysts. Finally, along the left side, the greater the percentage of appropriate indicators available to analysts, the more accurate the alerting and the higher the analyst performance will be. Closing the loop, this improved performance decreases the insider's opportunity to commit the crime. This feedback loop, shown in green and labeled *B1*, is self-balancing in nature because the purpose of insider threat detection is to stem the flow of insider attacks.

Unfortunately, as shown in the Supervisor Reporting pattern, immediate supervisors often do not sufficiently report inappropriate or suspicious employee behaviors. Figure 3 below extends the B1 self-balancing loop with two self-reinforcing loops that characterize the nature of the problem. The R1 feedback loop shown in orange characterizes the effect of the trust trap on supervisors. Supervisors' knowledge of insider indicators influences the risk they perceive due to the insider. Of course, other factors exist as well: They just may not be aware of the importance of the indicators, or they may just be overwhelmed with their day-to-day responsibilities. In any case, the perceived risk of insider threat influences supervisors' trust in the insider, which in turn influences the extent of supervisors' monitoring for insider indicators. It is exactly that monitoring that supports supervisors' knowledge of insider indicators and ultimate risk. As described in the Trust Trap Mitigation pattern, the problem occurs when supervisors' trust of an insider is inappropriately high, leading to decreased monitoring and perceived risk.

Figure 3 also illustrates, via the R2 self-reinforcing feedback loop shown in blue, that a lack of sufficient supervisor monitoring can also lead to lower perceived risk on the part of insiders. That perception can, in turn, increase their incentive (or decrease their disincentive) for committing the crime. This dynamic leads to a form of unobserved emboldening of the insider, whereby the malicious insider actions that are not observed or acted upon spur continuance or escalation of the crime.

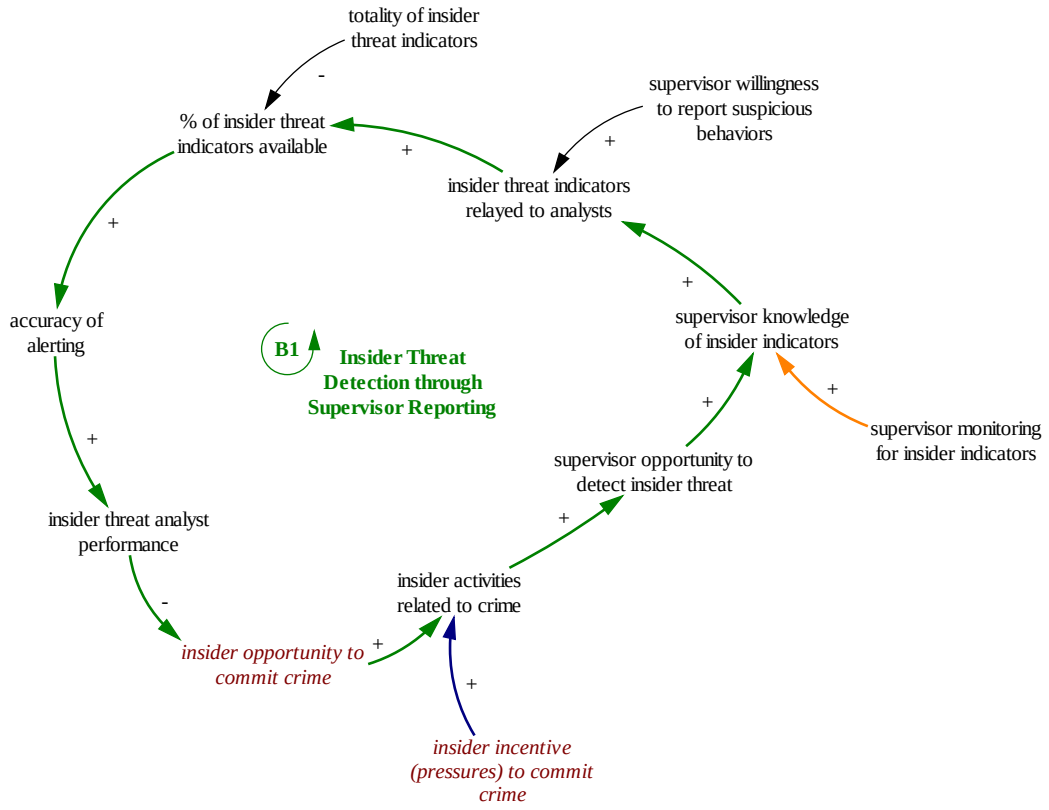

Figure 2: Insider Threat Detection Through Supervisor Reporting

Of course, inaccurate alerting or inappropriate insider investigation can lead to a host of other problems as well, including violations of privacy, or even worse, accusation and prosecution of innocent insiders. Organizations must always be careful to ensure they are monitoring employees according to applicable laws and industry norms. For example, organizations should properly obtain insiders' consent to monitoring as a condition of employment.

4 Foundations for Insider Threat Detection

Fundamental to the insider threat detection problem is the similarity of malicious activity to behaviors of nonmalicious insiders, perhaps performed as a normal part of their job. Signal detection theory provides important theoretical foundations for this problem (Swets, Dawes, & Monahan, 2000).

Figure 4 depicts the basic, *malicious insider* detection problem in terms of *risk score*.[9] The risk score is determined by an algorithm based on the weighted risk scores of a set of *rules* when applied to a particular individual. These rules are based on behavioral and technical indicators displayed by employees. Different rules may exist for different aspects of risk, such as risk related to an individual's travel activity or financial transactions. A composite risk score is based on the logical combination of multiple rules for an individual. The rules should be such that a

[9] For the purposes of this description, we assume that the risk scores of the insider populations are normally distributed.

large percentage of (nonmalicious) employees do not achieve a great enough risk score to cross the risk threshold.

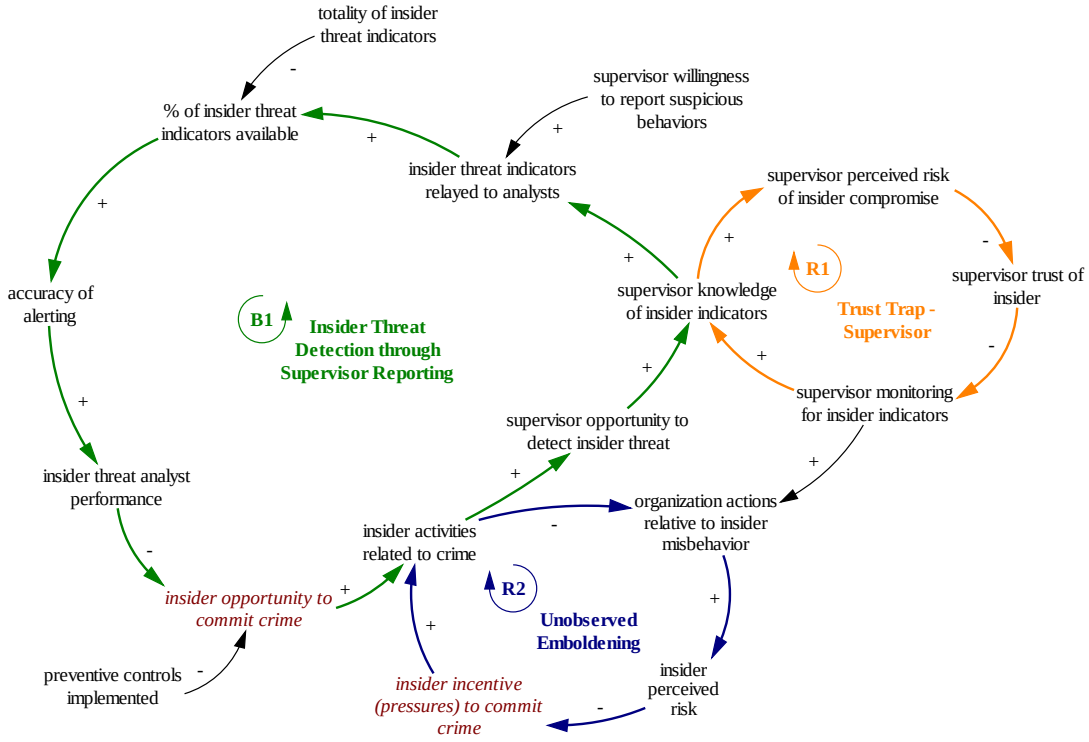

Figure 3: Trust Trap Influence on Supervisor Reporting

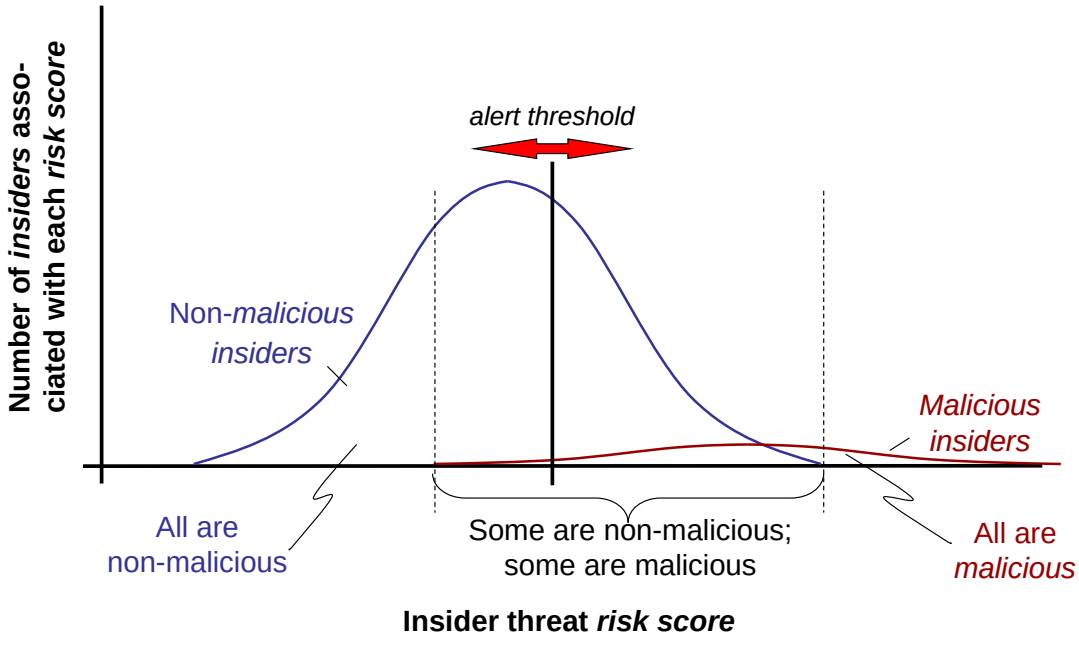

Figure 4: Insider Threat Detection Problem and Risk Threshold Setting

The overlap of the two curves, shown between the two dashed lines, represents those behaviors exhibited by both nonmalicious and malicious insiders. Managers must decide where in this range to set the *risk threshold*—the value of a risk score above which an individual is identified as suspicious (i.e., should be further investigated). Generally, greater risk thresholds catch fewer nonmalicious insiders but also fewer malicious insiders. Lower risk thresholds catch more malicious insiders but also more nonmalicious insiders.

Over time, managers can improve the formulas for risk scores. Ideally, this change would shift the population of malicious insiders to the right of the graph and the population of nonmalicious insiders to the left, decreasing the overlap. The risk scores fall into one of four categories: true positive, false positive, true negative, and false negative.

- A true positive occurs when an employee's risk score accurately identifies that employee as a malicious insider. True positives are valuable because they allow organizations to recognize and prevent malicious insider actions. In Figure 4, true positives are the malicious insiders whose risk scores fall above the risk threshold.
- A false positive occurs when a nonmalicious employee's risk score wrongly identifies that employee as a malicious insider. This costs the organization the time required to review the employee's activities and confirm them as nonmalicious. In Figure 4, false positives are the nonmalicious insiders whose risk scores fall above the risk threshold.
- True negatives occur when a nonmalicious employee's risk score falls below the risk threshold, correctly indicating that the employee is not a malicious insider. In Figure 4, true negatives are the nonmalicious insiders whose risk scores fall below the risk threshold.
- False negatives occur when a malicious insider's risk score falls below the risk threshold and the insider is incorrectly categorized as a nonmalicious employee. These can be the most costly errors to an organization because of the potential damage that an insider attack can cause. In Figure 4, false negatives are the malicious insiders whose risk scores fall below the risk threshold.

It is important that the insider threat detection program be improved continuously over time. Two aspects critical to improving the cost effectiveness of an insider threat detection program are the accuracy of the rules that create risk scores and the location of the risk threshold relative to the populations of malicious insiders and nonmalicious employees. The accuracy of the rules that create risk scores depends on understanding the true *indicators* of insider risk. As previously mentioned, improving the accuracy of the risk scores over time separates the distributions of nonmalicious and malicious insiders from each other. This lowers cost to the organization by reducing the number of false positives and false negative.

While adjusting risk score calculations can increase the accuracy of identifying malicious insiders, all organizations can expect some overlap of the risk scores of malicious and nonmalicious employees. Setting the appropriate risk threshold depends on understanding the importance of catching the malicious insiders before they cause harm versus the importance of not implicating nonmalicious insiders in malicious activity. Setting the risk threshold too high will allow malicious insiders to go undetected until it is too late. Setting the risk threshold too low will waste resources on fruitless investigations and incur human costs associated with wrongful implication of the innocent.

Figure 4 depicts the complexity associated with the insider threat detection problem due to the relative infrequency of insider incidents and the large overlap of the distributions of nonmalicious and malicious insiders. Large numbers of false positives can easily overwhelm analysts.

5 A Preliminary System Dynamics Model Foundation

To have an effective insider threat analysis, the organization must define, capture (or gather), analyze, and act upon indicators that define the risk score. These necessary steps must all be performed in a timely fashion. However, potentially important indicators of insider threat are frequently not

- identified as they happen
- recorded by organizational departments
- relayed to insider threat analysts
- recorded or relayed in a timely manner

This lack of proper indicator handling creates lag in an organization's awareness of increasing risk of insider compromise, which can inhibit effective insider threat defense. The model we are developing assumes that insider threat detection can be inhibited in this way. Appendix C depicts the full model.

Figure 5 shows a portion of the model with two similar segments: the bottom segment (3b) shows the processing of true positives indications (TPIs), and the top segment (3a) shows the processing of false positives indications (FPIs).[10] (Appendix A defines the notation used by system dynamics modeling.) Both true positive indications and false positive indications represent a series of events that have been flagged as a potential malicious insider attack. The security analyst's job can be viewed as distinguishing between these two. In true positive indications, the actions of a malicious insider caused the indicators to cross the risk threshold. False positive indications confuse the analyst's job by providing data that look like the actions of a malicious insider but are actually those of an employee performing his or her normal duties.

Analysis of TPIs and FPIs is necessary and valuable to the organization. TPIs allow organizations to recognize and mitigate malicious insider activity. TPIs may require simple review before becoming apparent or require a more thorough investigation. FPIs are valuable because they allow the organization to refine the risk threshold algorithm. Refining the risk threshold over time based on false positives and false negatives allows the organization to improve the accuracy of its detection program over time.

The model assumes there is only one attacker at a time.[11] The simulation starts at hiring time with some set of predisposition indicators (i.e., predisposition TPIs). Figure 5b shows the processing of TPIs generated by the malicious actions of an attacker. We distinguish between

[10] Defining insider activity as a false positive or a true positive requires the organization to have noticed the activity and classified (or misclassified) it as malicious activity. We assume that events classified as TPIs and FPIs may or may not be noticed by the organization. If they are noticed and acted on, then the individuals exhibiting TPIs and FPIs become true positives and false positives, respectively.

[11] While this is not always the case, it is a reasonable simplification for this preliminary model given the low base rate of insider incidents.

behavioral TPIs, which involve personal or interpersonal behaviors, and technical TPIs, which involve the use of information technology. The starting point of the attack—both the technical and behavioral aspects—and the duration of the attack are fully parameterized. Behavioral and technical TPIs may be generated (by the insider threat actions), recorded (by the responsible organizational departments), and relayed to analysts as shown along the perimeter of the TPI model segment. Of course, some TPIs may not be relayed to analysts or recorded.

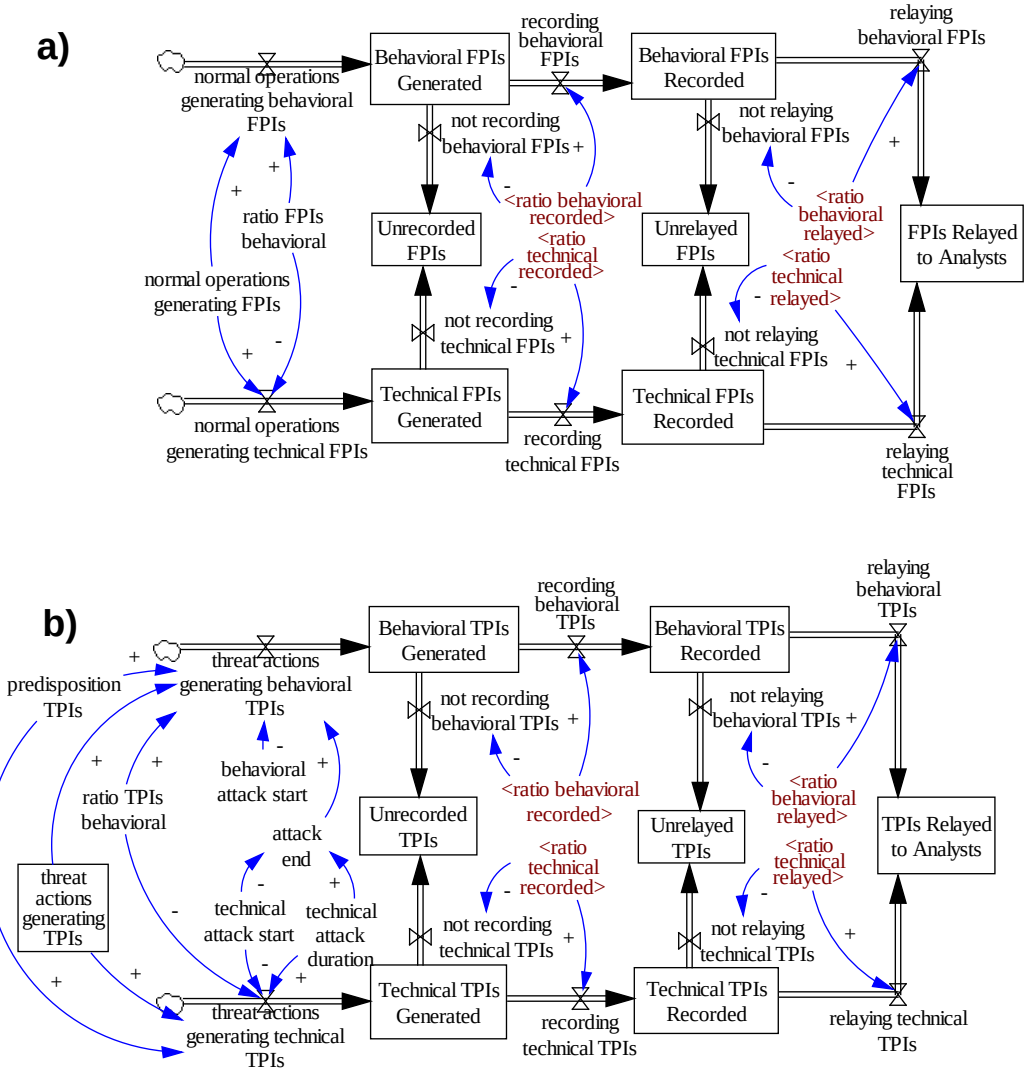

Figure 5: Generating, Recording, and Relaying (a) False Positive Indications (FPIs) and (b) True Positive Indications (TPIs)

The stocks of unrecorded or unrelayed TPIs are missed opportunities to account for indicators of increased insider risk. We assume that something not recorded is not remembered. Other variables in the TPI model segment represent the parameters of the model that can be instantiated once the baseline enterprise architecture is specified, for example, the ratio of

- behavioral TPIs (versus technical TPIs)
- technical/behavioral TPIs recorded (versus TPIs not recorded)
- technical/behavioral TPIs relayed to analysts (versus TPIs not relayed to analysts)

Figure 5a represents a very similar stock and flow structure for processing FPIs. Here, the parameters are generated from the properties of the normal (nonmalicious) population of insiders.

Figure 6 extends the stock and flow model of Figure 5: the variables “FPIs Relayed to Analysts” and “TPIs Relayed to Analysts” at the right end of Figure 5 are repeated at the left end of Figure 6. In this part of the model, there is much more interaction between the FPI and TPI segments than in Figure 5. This is because although FPIs and TPIs occur in separate stocks in the model, we do not assume that the analysts (i.e., the “Reviewers” and the “Investigators”) immediately know the difference. To them, the TPIs and FPIs are one big collection of useful information that they need to analyze to distinguish wheat from chaff. This model assumes a two-stage analysis of indicators: If an initial review by the “Reviewers” of the TPIs and FPIs is inconclusive, a full investigation is conducted by the “Investigators.” Further, the rates of incident review and incident investigation are the same for malicious and nonmalicious acts. Inconclusive analysis leads to stocks of “FPIs Unresolved” and “TPIs Unresolved” initially, but eventually all FPIs and TPIs are identified as such.

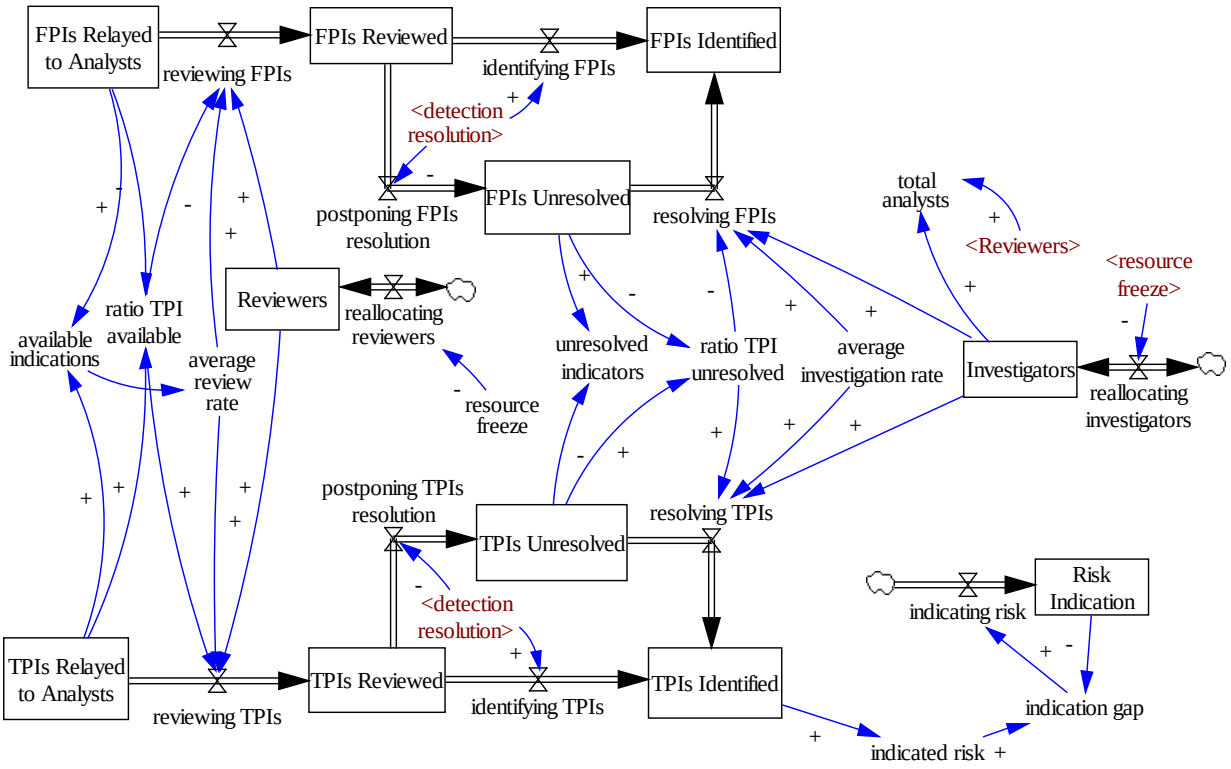

Figure 6: Reviewing, Identifying, and Resolving False Positive Indications (FPIs) and True Positive Indications (TPIs)

As shown in the lower right corner of Figure 6, the risk indication grows as more TPIs are identified. In this preliminary model, the indicators are all equal in terms of perceived risk, though earlier indicators weigh more than later ones.

6 Model Execution

Instantiating the parameters of the model described in the last section requires detailed information or estimates about the organization or type of organization in which the patterns are to be used. To make our approach as generic as possible, we specify a baseline enterprise architecture that represents a class of organizations our results will apply to.

That baseline is characterized by a set of available data sources, both human and technological, and the extent to which the indicators collected by those data sources are recorded and relayed to insider threat analysts. We need to specify the scale of the detection task for the organization, including the size of the workforce and the average number of malicious insiders the organization could have over some time period of interest. The organization needs to identify the following aspects of the behavioral and technical indicators:

- the fraction that are recorded (by anybody in the organization) and the average time it takes to record them
- the fraction that are relayed to insider threat analysts and the average time it takes to relay them

Measures involving the two-stage analysis of indicators also need to be estimated, such as the extent to which an in-depth investigation of indicators is needed and the rates of both initial review and further investigation.

Once we have the baseline enterprise architecture, we can use data derivable from the CERT insider threat database to measure aspects such as the range of data sources that would have been useful for detecting malicious insiders. Other parameters about the actual attack can also be derived, such as

- average start time of behavioral indicators (after hiring)
- average start time of technical indicators (after hiring)
- time lag of technical indicators behind behavioral indicators
- average duration of attack

We are currently conducting an analysis to determine such measures derived from the CERT database. However, to demonstrate the model's potential we did a quick analysis of 60 cases in the database and partitioned their indicators into behavioral and technical. An intuitive understanding of which indicators are likely to be recorded by the average organization and which are likely to be relayed to some type of investigator led to the following breakdown:

- fraction behavioral 0.74
- fraction technical 0.26
- fraction behavioral recorded 0.64
- fraction technical recorded 0.96
- fraction behavioral relayed 0.09
- fraction technical relayed 0.88

We used an abstract interface to the simulation model (shown in Appendix D) to test the model's behavior under a range of settings. Figure 7 shows the output of the model simulation based on the above parameters for two simulation runs. The first run, "Low Beh Relayed," uses the 0.09 fraction of the total insider behavioral indicators relayed to analysts. The second run, "High Beh Relayed," uses 10 times the "Low Beh Relayed" fraction (i.e., 0.9) as the fraction of behavioral indicators relayed. All the other parameters remained the same. As shown, for a risk threshold of

.025 in the range 0 to 1, the detection time drops from about 63 weeks to about 33 weeks. Appendix E shows the distribution of nonmalicious and malicious populations used for this sample.

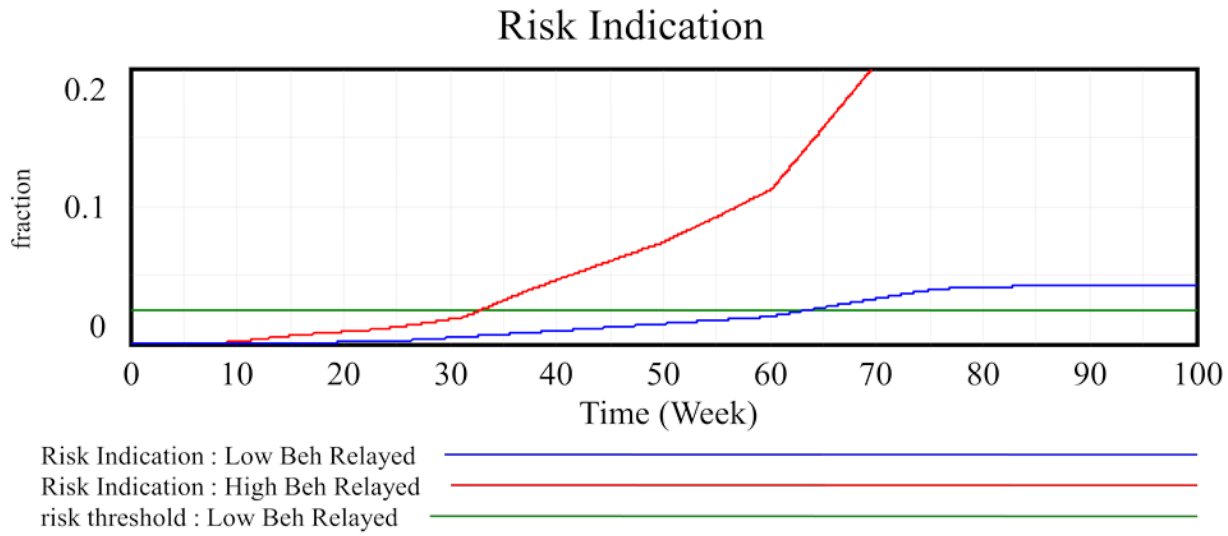

Figure 7: Risk Indication for Sample Data

7 A Proposed Study to Collect Data as Input to the Model

In many organizations, the responsibility for managing the insider threat falls almost exclusively with the information technology staff. But many of the early indications of problems on the horizon occur at a behavioral, nontechnical level. The aim of this study is to determine how much earlier the detection of increased insider threat risk can be advanced by considering concerning behavioral (not necessarily visible online) *observables* of insider IT sabotage in addition to concerning technical (visible online using common network/system logging) observables. Of all the classes of insider incidents, we chose insider IT sabotage because we believe its behavioral observables are more likely to indicate true malicious behavior.

The timing of the detection is a critical input to our simulation model because it determines how soon behavioral indicators arise prior to the technical indicators. If the detection is early enough (say by at least a few days), organizations monitoring behavioral as well as technical observables may be able to prevent insider attacks, as opposed to cleaning up the damage after an attack. At the very least, earlier detection will allow the organization to better mitigate the attack, possibly limiting its damage.

Past studies suggest a temporal precedence relationship between behavioral and technical observables in insider incidents (Shaw & Fischer, 2005) (Keeney et al., 2005) (Moore, Cappelli, & Trzeciak, 2008). But these studies are not explicit about the exact nature of that temporal relationship or even about how often these observables would occur in the workplace as true or false positives.

We assume that the observables are likely to occur quite often even without an associated insider attack. This would make false positives problematic for most organizations' insider threat programs. However, we believe that the incidence of both concerning behavioral and technical

observables within a year of each other will be much lower for employees not engaging in malicious activity. We base the length of this one-year period on unpublished indications that the timeframe of insider IT sabotage almost always evolves within the year prior to the initial damage from the incident. This forms the basis for our first hypothesis.

- Hypothesis 1: Insider IT saboteurs exhibit both concerning behavioral and technical observables more often than nonsaboteurs do. For our purposes we are interested only in concerning behavioral observables that occur within a year of concerning technical observables.

Our second hypothesis makes explicit the temporal relationship between behavioral and technical observables in insider IT sabotage attack.

- Hypothesis 2: In the timeline of insider IT sabotage attacks, the first concerning behavioral observable occurs before the first concerning technical observable.

If concerning behavioral observables occur significantly before the concerning technical observables, organizations monitoring for both can detect insider IT sabotage significantly earlier than organizations performing technical monitoring only.

7.1 Approach

A database of previously collected cases of insider IT sabotage will provide the population of malicious actors (Cappelli, Moore, & Trzeciak, The CERT Guide to Insider Threats, 2012, p. 325). As described in the Multiple Case Study Methodology proposed by (Yin, 2009), case studies should be selected as a laboratory manager selects the topic of a new experiment: to test specific hypotheses that will help to validate, extend, or modify existing theory. Selecting multiple cases is analogous to replicating experiments to gain further confirming evidence or to test the limits of the experimental results. We intend to analyze the occurrence and timing of the first concerning technical and behavioral observables in the selected cases.

A comparison group is also needed to determine how often concerning behavioral and technical observables occur for nonmalicious employees, in other words, employees who do not eventually engage in insider IT sabotage incidents. To increase the value of the comparison of the two groups, the comparison group will be matched with the experimental group based on key attributes.

We will be assessing insider IT saboteurs from the experimental group and the employees from the comparison group to determine how often each group had at least one concerning behavioral observable and one concerning technical observable. Insider IT saboteurs having a statistically greater prevalence would suggest that having both types of observables is a good indicator of heightened insider IT sabotage risk.

The next question is how much earlier, if at all, is the detection of behavioral observables (N_{beh}) versus the detection of technical observables (N_{tech}). As shown in Figure 8, we measure this notice as the distance in time from the initial point of observable sabotage damage. The second hypothesis tests whether N_{beh} is greater than N_{tech}.

If this hypothesis holds, we will analyze the distribution of $N_{beh} - N_{tech}$ because this will indicate how much before the first technical observable the first nontechnical observable occurs. If the

mean length of the behavioral notice prior to technical notice is sufficient for the staff to prepare additional monitoring activity, the value of the notice from behavioral observables is enhanced.

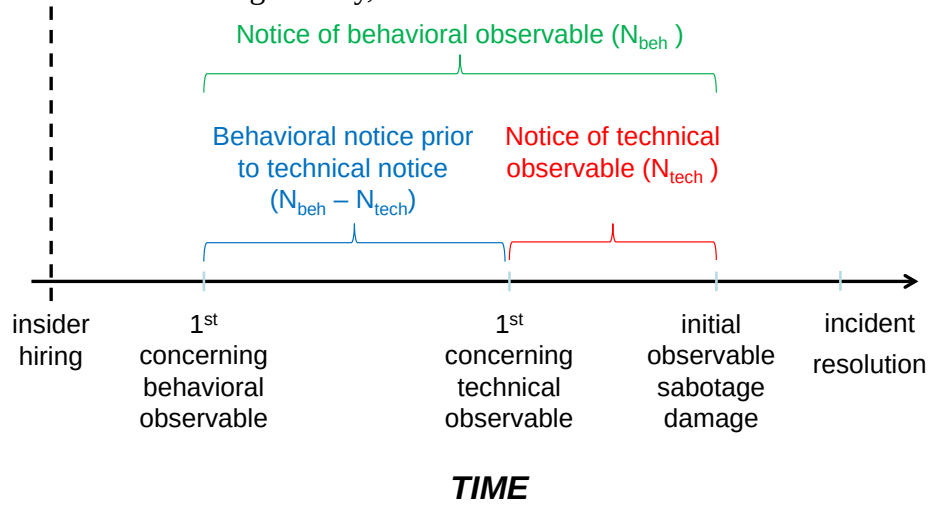

Figure 8: Key Measures Driving Data Collection

7.2 Expected Results

The proposed study will provide critical input data to our model. Also, provided that the hypotheses are supported, the proposed study should motivate and justify an organization's transition to a comprehensive and fully integrated insider threat program. Such a program would include individuals from across the organization's departments, especially human resources and information technology, and improved communication with managers from other departments. Tearing down an organization's stovepipes to make this happen will be challenging. Cross-department representation in an insider threat incident management team will help to build bridges. Sharing data across teams may be the most challenging task and may encounter cultural, policy-related, and legal hurdles. However, we believe a more integrated strategy to tackle this problem is worth the effort to lower operational risk and improve efficiency.

8 Conclusion

This report describes a modeling and simulation foundation, based on the system dynamics methodology, to test the efficacy of these insider threat detection controls prior to pilot testing. This paper describes the first stage of our overall effort. In addition to collecting more data to ground our model, as described in section 7, we plan to form partnerships with organizations that have active insider incident investigation teams. The parameters of the model will then be able to be specified based on the operational profile of the organization's business processes, incident investigation approach, and insider threat history. This will permit testing of insider threat detection approaches virtually within the organization and making recommendations accordingly for pilot testing those approaches.

Ultimately, we hope this work will provide organizations an approach for making strategic decisions to mitigate the insider threat. The approach will gain credibility from its use of established theories in related areas and the scientific approach of using simulation models to test key hypotheses prior to pilot testing. This work should improve enterprise, system, and software

architecture in a way that operationally reduces both the number and impact of insider attacks on an organization's information assets.

9 Acknowledgements

We would like to thank editors Paul Ruggiero and Pennie Walters for their excellent suggestions for improving this paper.

This material is based upon work funded and supported by the Department of Defense under Contract No. FA8721-05-C-0003 with Carnegie Mellon University for the operation of the Software Engineering Institute, a federally funded research and development center.

DM-0000143

References

Blackwell, C. (2009). A Security Architecture to Protect against the Insider Threat from Damage, Fraud, and Theft. *Proceedings of the 5th Annual Workshop on Cyber Security and Information.* Knoxville, TN: ACM.

Bulling, D., Scalora, M., Borum, R., Panuzio, J., & Donica, A. (2008). *Behavioral Science Guidelines for Assessing Insider Threats.* Lincoln: Public Policy Center, University of Nebraska. Retrieved from http://digitalcommons.unl.edu/publicpolicypublications/37

Cappelli, D. M., Moore, A. P., & Trzeciak, R. F. (2012). *The CERT Guide to Insider Threats: How to Prevent, Detect, and Respond to Information Technology Crimes (Theft, Sabotage, Fraud).* Addison-Wesley.

Caputo, D. D., Stephens, G. D., & Maloof, M. A. (2009, November/December). Detecting Insider Theft of Trade Secrets. *IEEE Security and Privacy,* 14-21.

CSO Magazine, U.S. Secret Service, SEI CERT, Deloitte. (2011, January). 2011 CyberSecurity Watch Survey: Organizations Need More Skilled Cyber Professionals to Stay Secure. *CSO Magazine*. Retrieved from http://www.cert.org/archive/pdf/CyberSecuritySurvey2011.pdf

DARPA Information Innovation Office. (2010). *Anomaly Detection at Multiple Scales (ADAMS): Broad Agency Announcement (DARPA-BAA-11-04).* Arlington, VA: DARPA.

DARPA Strategic Technology Office. (2010). *Cyber Insider Threat (CINDER) Broad Agency Announcement (DARPA-BAA 10-84).* Arlington, VA: DARPA. Retrieved from https://www.fbo.gov/utils/view?id=16a7be70c7ef8c965da695fe8f0ecb50

Duran, F., Conrad, S. H., Conrad, G. N., Duggan, D. P., & Held, E. B. (2009, November/December). Building a System for Insider Security. *IEEE Security and Privacy*, 30-38.

Executive Office of the President. (2011). *Memorandum M-11-08, Initial Assessments of Safeguarding and Counterintelligence Postures for Classified Nationall Security Information in Automated Systems.* Washington, DC.

Fischer, L. (2003). *Characterizing Information Systems Insider Offenders.* Pensacola, FL: International Military Testing Association Proceedings.

Greitzer, F. L., Paulson, P. R., Kangas, L. J., Franklin, L. R., Edgar, T. W., & Frinke, D. A. (2009). *Predictive Modeling for Insider Threat Mitigation.* Department of Energy. Retrieved from http://www.pnl.gov/cogInformatics/media/pdf/TR-PACMAN-65204.pdf

Ingram, R. E., & Price, J. M. (2001). *Vulnerability to psychopathology: Risks across the lifespan.* New York: Guilford.

Keeney, M., Kowalski, E., Cappelli, D. M., Moore, A. P., Shimeall, T. J., & Rogers, S. (2005). *Insider Threat Study: Computer System Sabotage in Critical Infrastructure Sectors.* Pittsburgh, PA: Software Engineering Institute and United States Secret Service. Retrieved from http://www.cert.org/archive/pdf/insidercross051105.pdf

Maybury, M., Chase, P., Cheikes, B., Brackney, Brackney, D., Matzner, S., . . . Lewandowski, S. (2005). Analysis and Detection of Malicious Insiders. *International Conference on Intelligence Analysis.* McLean, Va. Retrieved from http://www.mitre.org/work/tech_papers/tech_papers_05/05_0207/index.html

Meadows, D. (2008). *Thinking in Systems: A Primer.* Chelsea Green Publishing.

Monroe, S., & Simons, A. (1991). Diathesis-stress theories in the context of life-stress research: Implications for the depressive disorders. *Psychological Bulletin, 110*, 406-425.

Moore, A. P., Cappelli, D. M., & Trzeciak, R. F. (2008). The 'Big Picture' of Insider IT Sabotage Across U.S. Critical Infrastructures. In S. Stolfo, S. M. Bellovin, S. Hershkop, A. Keromytis, & S. Sinclair, *Insider Attack and Cyber Security: Beyond the Hacker* (pp. 17-52). Pittsburgh, PA: Carnegie Mellon University. Retrieved from http://www.sei.cmu.edu/library/abstracts/reports/08tr009.cfm

Predd, J., Pfleeger, S., Hunker, J., & Bulford, C. (2008, July/August). Insiders Behaving Badly. *IEEE Security and Privacy, 6*(4), 66-70.

Shaw, E., & Fischer, L. G. (2005). *Ten Tales of Betrayal: The Threat to Corporate Infrastructure by Information Technology Insiders Analysis and Observations.* Monterrey, CA: PERSEREC Technical Report 05-13. Retrieved from http://www.dhra.mil/perserec/reports/tr05-13.pdf

Software Engineering Institute. (2011). *2011 CyberSecurity Watch Survey.* Pittsburgh: Carnegie Mellon University. Retrieved from http://www.cert.org/archive/pdf/CyberSecuritySurvey2011Data.pdf

Sterman, J. D. (2000). *Business Dyamics: Systems Thinking and Modeling for a Complex World.* McGraw-Hill.

Straub, D. W. (1986). *Controlling Computer Abuse: An Empirical Study of Effective Security Countermeasures.* Bloomington, IN: Indiana University.

Swets, J. A., Dawes, R. M., & Monahan, J. (2000). Better Decisions Through Science. *Scientific American, 283*(4), 82-87.

Willison, R., & Siponen, M. (2009, September). Overcoming the insider: reducing employee computer crime through Situational Crime Prevention. *Communications of the ACM, 52*(9), 133-137.

Yin, R. (2009). *Case Study Research: Design and Methods, 4th edition.* Sage Publications.

Appendix A: System Dynamics Modeling Notation

Figure 9 summarizes the notation used by system dynamics modeling. The primary elements are variables of interest, stocks (which represent collection points of resources), and flows (which represent the transition of resources between stocks). Signed arrows represent causal relationships, where the sign indicates how the variable at the arrow's source influences the variable at the arrow's target. A positive (+) influence indicates that the values of the variables move in the same direction, and a negative (–) influence indicates that they move in opposite directions. A connected group of variables, stocks, and flows can create a path that is referred to as a feedback loop. System dynamics models identify two types of feedback loops: balancing and reinforcing. The type of feedback loop is determined by counting the number of negative influences along the path of the loop. An odd number of negative influences indicates a balancing loop; an even (or zero) number of negative influences indicates a reinforcing loop.

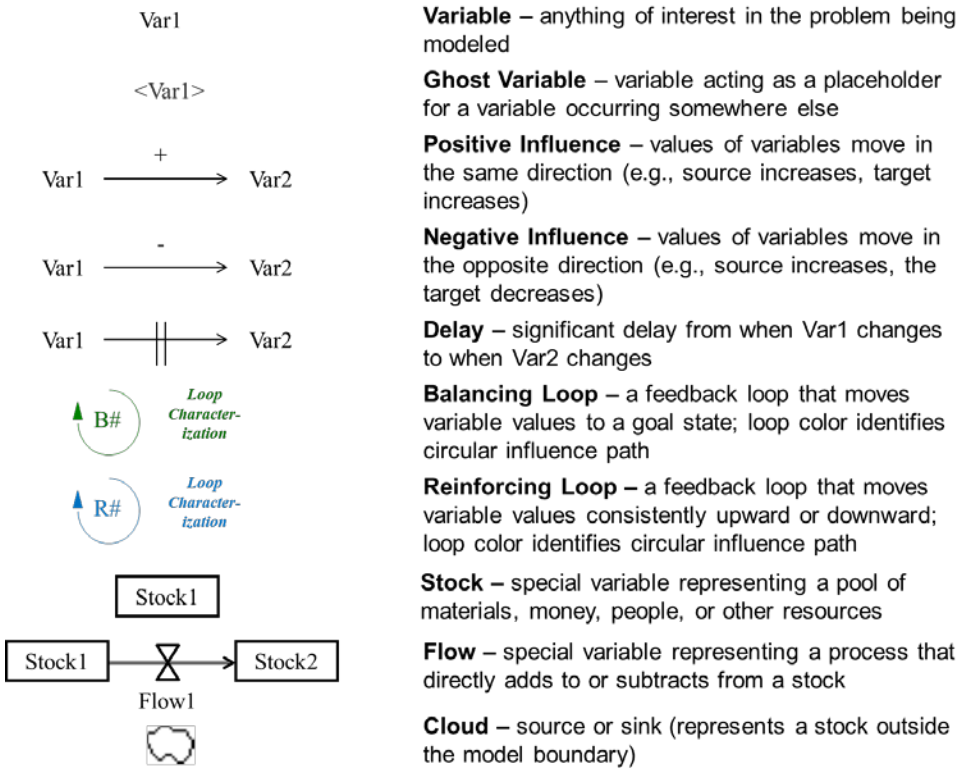

Figure 9: System Dynamics Notation

Significant feedback loops identified within a model are indicated by a loop symbol and a loop name in italics. Balancing loops—indicated with the label *B* followed by a number in the loop symbol—describe aspects of the system that oppose change, seeking to drive variables to some goal state. Balancing loops often represent actions that an organization takes to mitigate a problem. Reinforcing loops—indicated with the label *R* followed by a number in the loop symbol—describe system aspects that tend to drive variable values consistently upward or downward. Reinforcing loops often represent the escalation of problems but may include problem mitigation behaviors.

Appendix B: Glossary

behavioral
Involves personal or interpersonal behaviors.

behavioral observable
A behavioral action, event, or condition. We are generally interested in behavioral observables that are concerning, for example, intoxication during working hours.

false negative indication (FNI)
An incorrect indication of malicious activity as nonmalicious.

false positive indication (FPI)
An incorrect indication of nonmalicious activity as malicious.

false positive probability
The probability that a nonmalicious insider behavior is identified as suspicious (incorrect prediction).

indicator
A set of observables that, in combination, indicates an increased malicious insider risk, for example, centralization of programs on a central server combined with attempts to undermine recovery and backup processes.

insider
A current or former employee, contractor, or other business partner of an organization.

malicious insider
An insider who engages in malicious insider activity; malicious insiders include spies, fraudsters, information thieves, and saboteurs.

malicious insider activity
Activity associated with insider incidents of interest.

malicious insider risk
The potential for harm due to malicious insider activity.

observable
An (individual or organization) action, event, or condition that could be observed from a detection source.

online detection source
A source of data available electronically for detecting observables.

precursor
An action, event, or condition that precedes insider incidents of interest and is hypothesized to be associated with those incidents. Precursors that can be observed and definitely linked to malicious insider activities are indicators of increased malicious insider risk.

risk score
A measure of some aspect of malicious insider risk. Different rules or rule sets may exist for different aspects of risk, for example, risk based on an individual's travel activity or financial transactions. A composite risk score might be based on the logical combination of multiple rule sets for an individual.

risk threshold
A value of a risk score above which an individual is identified as suspicious (i.e., should be further investigated).

rule
A mapping from a set of indicators to a risk score for a particular individual; for example, an individual's centralization of programs on a central server combined with the insider's attempt to undermine recovery and backup processes indicates high risk.

rule set
A set of related rules and an algorithm that together return a risk score based on the weighted risk scores of the component rules when applied to a particular individual.

technical
Involves the use of IT.

technical observable
A technical action, event, or condition. We are generally interested in technical observables that are concerning, for example, an account audit reveals an unauthorized account.

true negative indication (TNI)
The correct indication of nonmalicious activity as nonmalicious.

true positive indication (TPI)
The correct indication of malicious activity as malicious.

true positive probability
The probability that a malicious insider behavior is identified as suspicious (correct prediction).

Appendix C: Insider Threat Detection Model

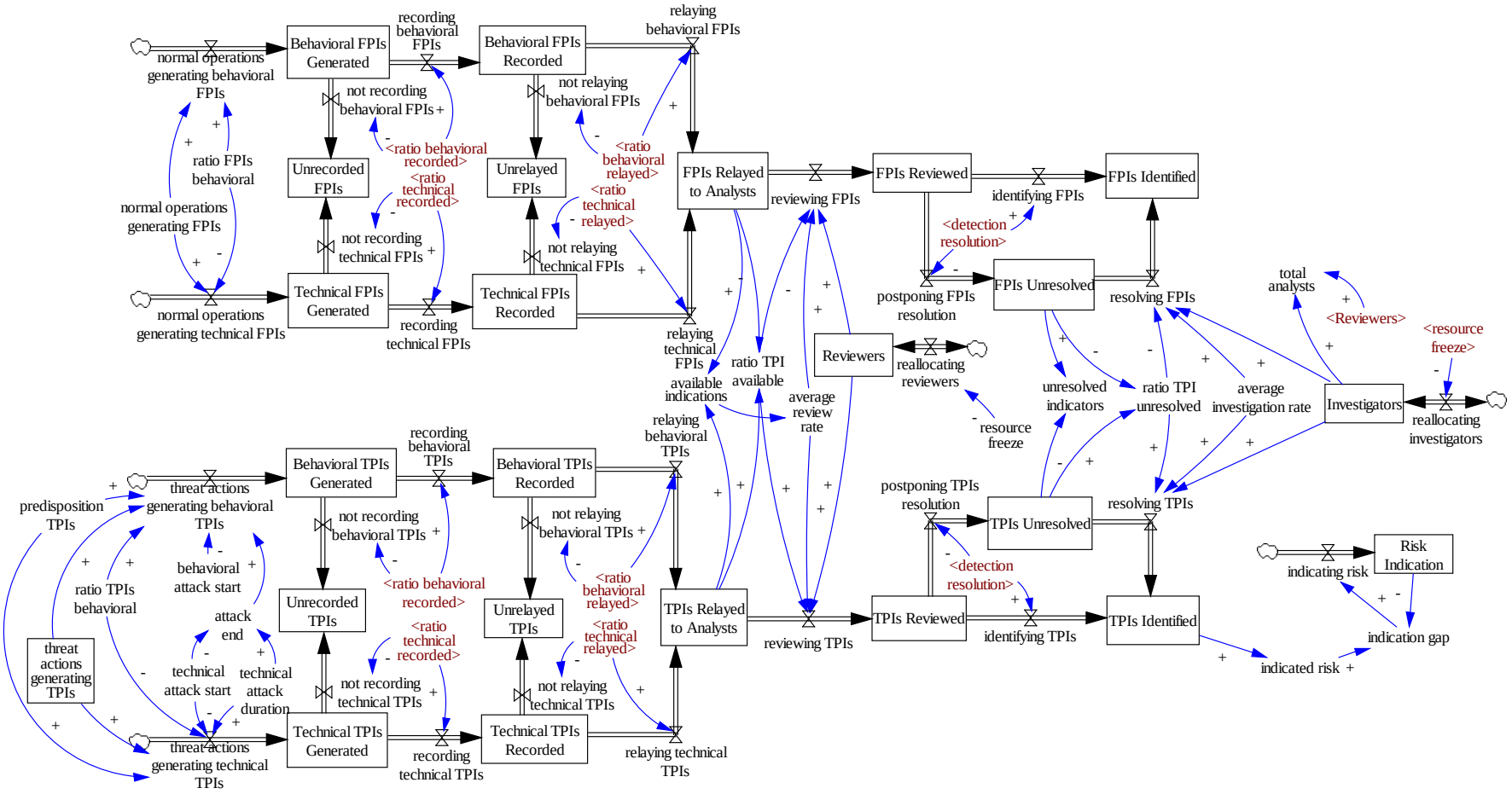

Figure 10: Insider Threat Detection Model

Appendix D: Simulation Model Interface

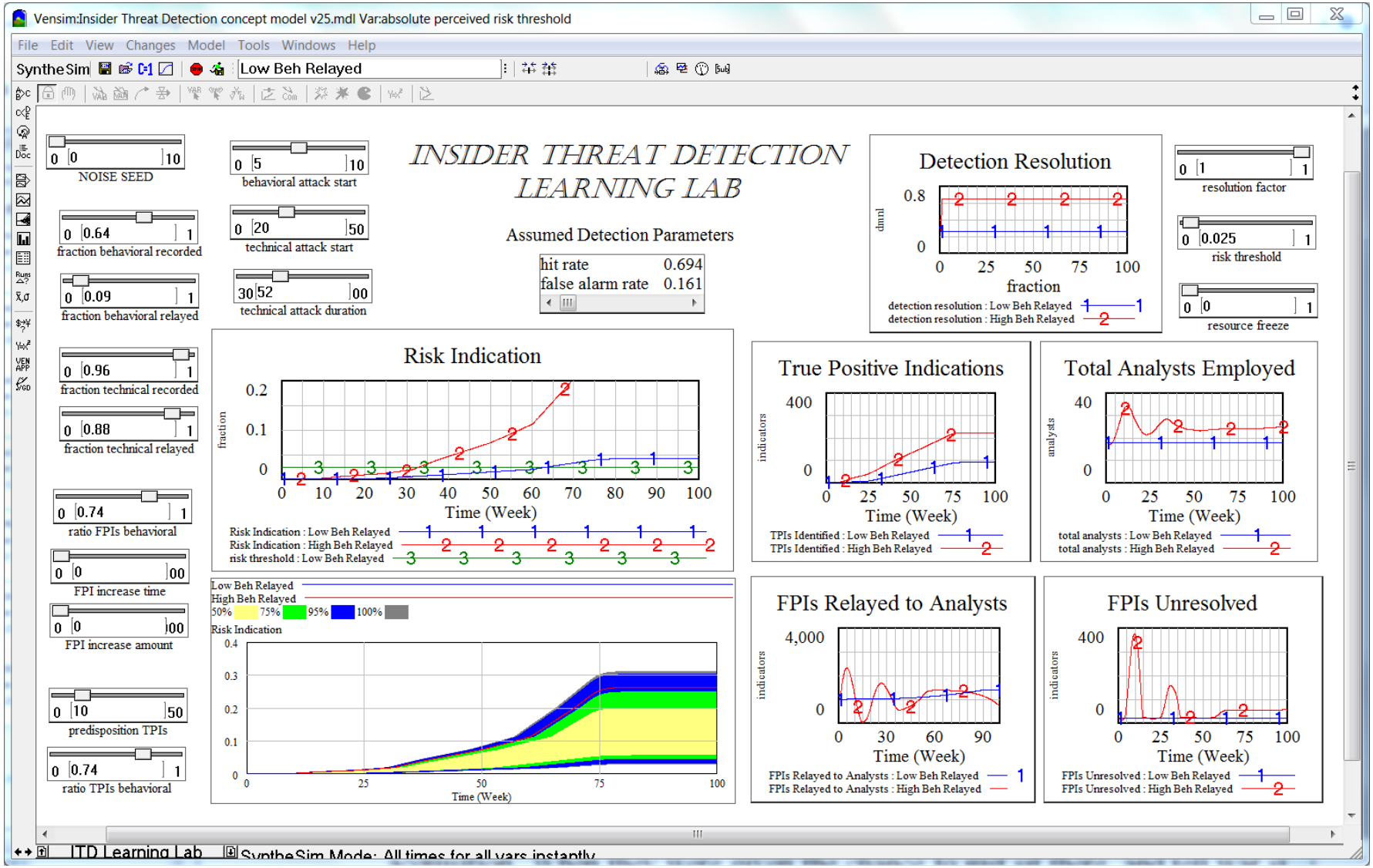

Figure 11: Simulation Model Interface

Appendix E: Assumed Distributions for Sample Analysis

We assumed normally distributed populations of nonmalicious (normal) and malicious populations with respect to their indication of risk. The figures below show the assumed distributions.

min perceived risk of normal	0	Total Signal	999,300
max perceived risk of normal	400	total number of FPIs	161,671
mean perceived risk of normal	200	false alarm rate	0.1617
std dev of perceived risk of threat	35		

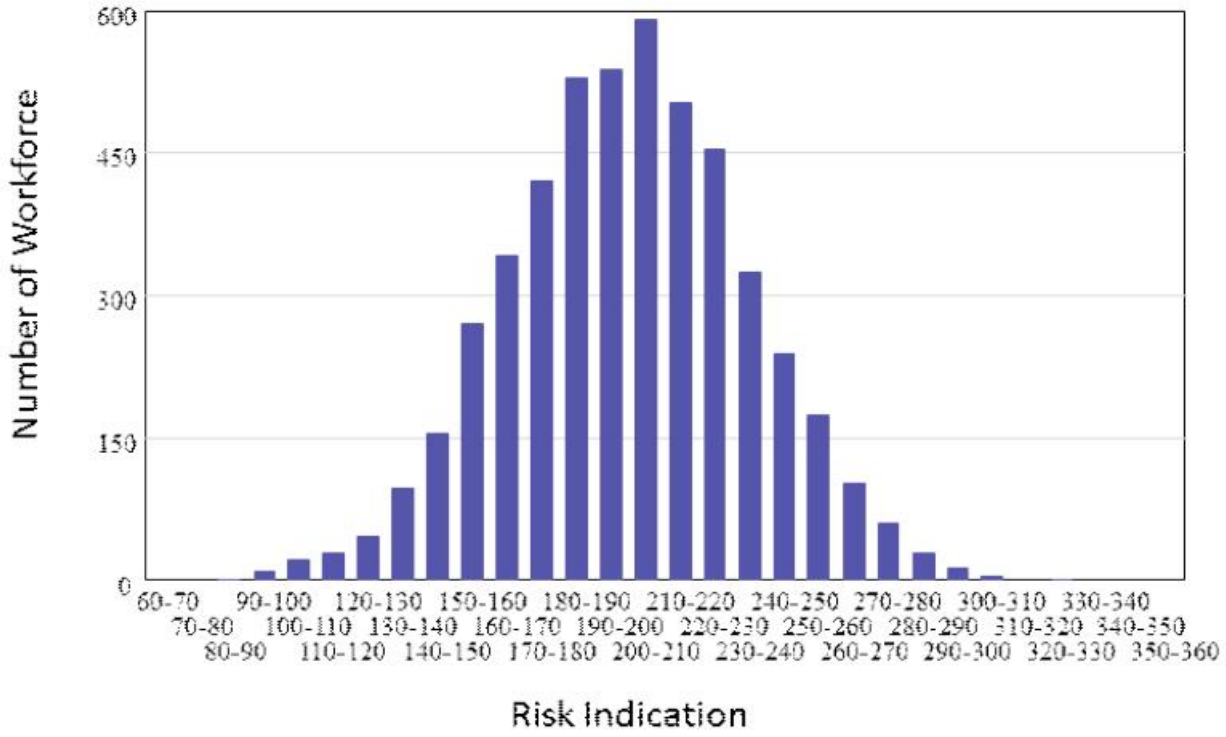

Figure 12: Nonmalicious Insider Population

min perceived risk of threat	175	Total Signal	1,460
max perceived risk of threat	500	total number of TPIs	1,014
mean perceived risk of threat	300	hit rate	0.6947
std dev of perceived risk of threat	70		

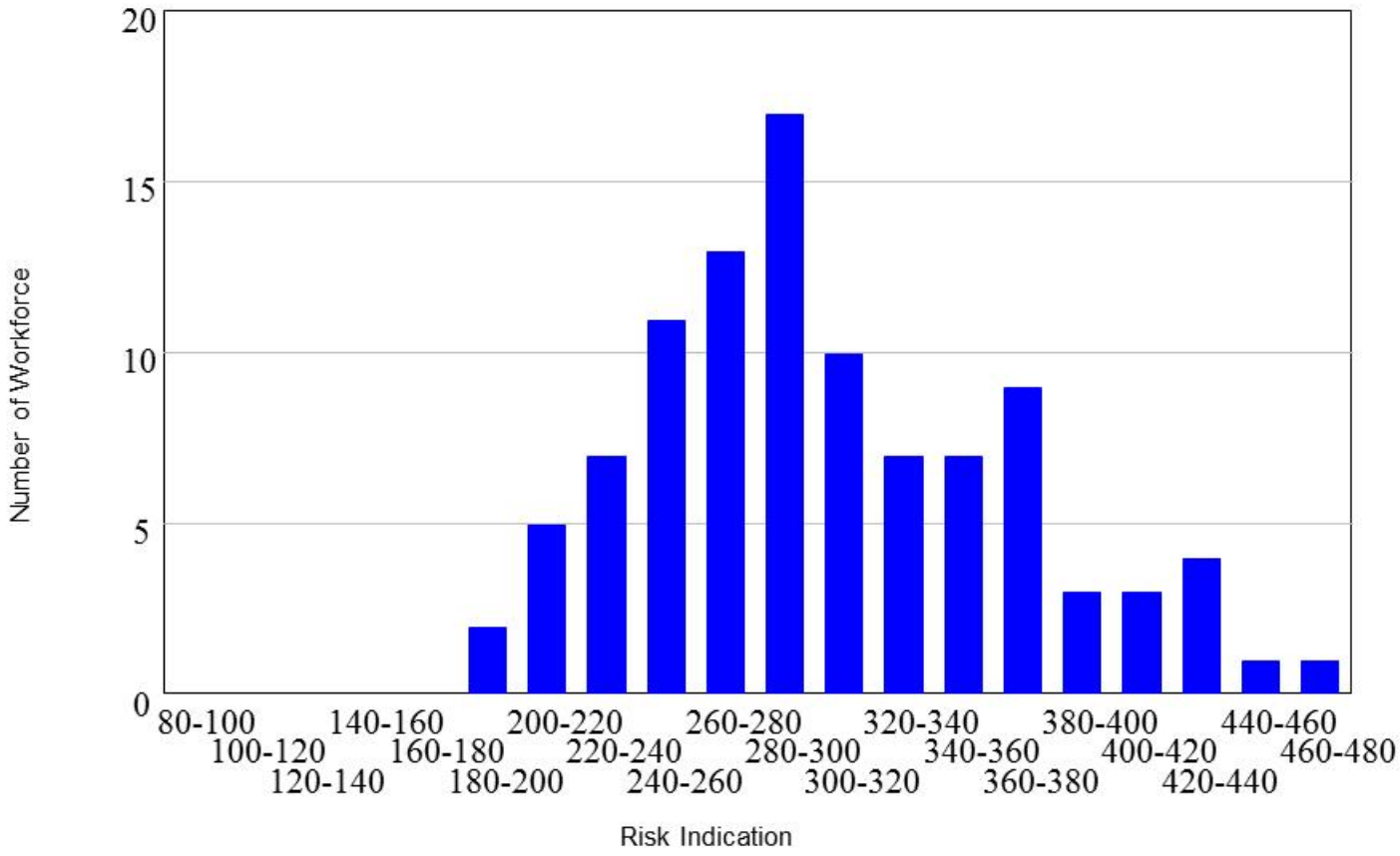

Figure 13: Malicious Insider Population