PNNL-SA-65204

Predictive Modeling for Insider Threat Mitigation

Frank L. Greitzer
Patrick R. Paulson
Lars J. Kangas
Lyndsey R. Franklin
Thomas W. Edgar
Deborah A. Frincke

April 2009

DISCLAIMER

PACIFIC NORTHWEST NATIONAL LABORATORY
operated by
BATTELLE
for the
UNITED STATES DEPARTMENT OF ENERGY
under Contract DE-AC05-76RL01830

Printed in the United States of America
Available to DOE and DOE contractors from the
Office of Scientific and Technical Information,
P.O. Box 62, Oak Ridge, TN 37831-0062;
ph: (865) 576-8401
fax: (865) 576-5728
email: reports@adonis.osti.gov

Available to the public from the National Technical Information Service,
U.S. Department of Commerce, 5285 Port Royal Rd., Springfield, VA 22161
ph: (800) 553-6847
fax: (703) 605-6900
email: orders@ntis.fedworld.gov
online ordering: http://www.ntis.gov/ordering.htm

Contents

This is a technical report on work conducted as Laboratory Directed Research and Develolpment, under the Pacific Northwest National Laboratory's *Information and Infrastructure Integrity Inititative*.

Introduction

Imagine a trusted system administrator, stressed and angry after going through a major life change and frustrated with unmet expectations at work. Approached by a rival firm, he is offered a job if he comes with "special expertise." Early indicators of such potential malicious insiders are present in both traditional and behavioral perspectives. Can these early warning signs be exploited by a monitoring system to prevent harm to the organization? How should the effectiveness of an automated insider threat tool be assessed? Intervention may ameliorate the problem or exacerbate a precarious situation. Would early intervention violate employee trust or legal guidelines? What about the potential for false accusations, and how these might affect an employee's career?

Scenario

Fred is a system administrator with a history of company loyalty and no performance issues. Lately, life seems to be treating him unfairly: he has marital and financial problems, and he is passed over for a job promotion he felt he deserved.

Frustrated, he argues with his manager, storms out of the office, and complains to Human Resources that the manager had been unfair. Fred subsequently has a similar heated exchange with his manager's boss. Months later, Fred is quietly "stewing" and recalls that a rival company offered him a position if he obtains guarded company secrets. Thinking his employer owes him, he begins to spend considerable time on the rival company's website and attempts to access his own company's protected data. He finds ways to "masquerade" as other colleagues and works odd hours to use their workstations; later he installs specialized scripts to download proprietary files automatically. He accesses his personal email account to collect and exfiltrate the downloaded files and covers his tracks by deleting the files from the computers he used. Then he makes plans to contact the rival company.

This paper describes a research framework designed to test the hypothesis that in cases of insider threat, predictive capabilities are enhanced by integrating employee data, i.e., psychosocial data, with the traditional cybersecurity audit data normally used by cyber analysts. Some questions that are addressed include: What data are available and effective as precursors or indicators of potential insider exploits? Which data are useful? How can predictive insider threat models be tested? What are associated practical and ethical issues in the work environment when an organization uses employee data to predict sabotage or espionage?

The Approach: Prediction with Messy Data

We restrict our consideration of "insiders" to members of an organization authorized to access its information system, data, or network with a degree of trust by the organization and who accept a commensurate level of scrutiny by the organization to deter possible abuse of these privileges. Recent studies (Keeney *et al*., 2005) and surveys (e-*Crime Watch Surveys,* available from the CERT web site, http://www.cert.org/insider_threat/) of cybercrime in both government and commercial sectors, reveal that current or former employees and contractors are the second greatest cybersecurity threat, exceeded only by hackers. The 2007 e-Crime survey showed that most insiders targeted proprietary information, including intellectual property, and customer and financial information. It has been argued that most threats could be prevented by "timely and effective action to address the anger, pain, anxiety, or psychological impairment of perpetrators who exhibit signs of vulnerability or risk well in advance of the crime of abuse." (Shaw & Fischer, 2005) This suggests that research is needed on predictive indicators and algorithms. However, despite considerable research into the psychology and motivation of insiders in the financial sector, it remains difficult to predict who will commit security fraud (Kramer *et al*., 2005). Potential benefits must be weighed against the possible adverse effect that an insider detection and mitigation strategy might have; certain policies, interventions, and monitoring might reduce the likelihood of an insider action in the first place, while others might add to a climate that breeds insider abuse.

Currently, no single threat assessment technique gives a complete picture of the insider threat problem. The IATAC SOAR report (Gabrielson *et al*., 2008) provides a comprehensive review. Typical approaches incorporate forensic measures including external threat/defense-oriented appliances such as Intrusion Detection or Prevention Systems (IDS/IPS). Research (see *Research on Psychology of Insiders*) suggests that a proactive approach must recognize possible precursors to malicious insider threat behavior that are manifested in employee stress, disgruntlement, and other signs.

Behavioral Monitoring Issues

Should behavioral warning signs be addressed proactively to prevent harm to the organization? Numerous studies have sought to identify the psychological profiles consistent with insider threat. These form a compelling body of research to warrant continued efforts to incorporate psychosocial factors into predictive models—see *Research on Psychology of Insiders*. The approach adopted in the research described in this paper focuses on predicting individual employee behaviors. Other approaches include models that describe the broader influences of management decisions, policies, and the work environment on employee behavior, which may be implemented using system dynamics models. (Moore *et al*., 2008).

Would the collection of psychosocial data violate employee trust or legal guidelines?

There is an inherent tension between an organization safeguarding its assets and employee privacy rights. The American Civil Liberties Union claims that "Electronic surveillance in the workplace is a major threat to [one's] right to privacy" (http://www.aclu.org/privacy/workplace/15646res20031022.html). Although law prohibits the sharing of personal information outside an organization, there are few restrictions on an employer's right to share it internally. (Lane, 2006, p. 261) Legal precedents uphold the employer's right to monitor employee data (emails, Internet use, etc.), particularly when employers provide notice to employees that their computers are subject to monitoring—In such cases there is "no reasonable expectation of privacy." At the heart of this question is the potential violation of employee

trust, privacy, or legal guidelines—see text box on *Privacy and Ethical Concerns*. Space does not permit a full treatment of the issues, but we suggest that while employers may monitor employees' cyber activity, there is a responsibility for safeguarding privacy, disclosure, and fairness: i.e., disclosure to employees about what is monitored and fairness in the sense that the process is applied uniformly.

Privacy and Ethical Concerns

It is widely acknowledged that employers have the right to monitor employee cyber activities and to internally share personal employee data (www.salon.com/tech/feature/1999/12/08/email_monitoring/print.html), although federal agencies can only use this information in certain circumstances (http://www.usdoj.gov/oip/privstat.htm). However, the ramifications of such use of information in terms of employee job satisfaction and public relations can be severe. Trust is a fundamental concept underlying the issue of privacy and workplace monitoring. Tabak and Smith (2005) assert that the initiation of trust and subsequent trust formation affects managerial implementation of electronic monitoring policies, and these policies have implications for workplace privacy rights. Similarly, employee perception of management practices influences employee trust in and commitment to the organization.

The privacy and ethics debate is clearly a contentious issue that deserves more discussion. (Greitzer & Endicott-Popovsky, 2008, took a first step in a forum discussion on security and privacy.) There is a fine line between what the organization "needs to know" and what is firmly in the realm of the employee's expectation of privacy. Indeed, a small percentage of employees actually engage in activities that would constitute "insider threat;" the rest of the population comprises honest, hard-working staff who would be highly offended to learn they were monitored. To the employer, the cost and damage of one instance of sabotage or espionage may warrant monitoring all behavioral and demographic employee data available to proactively prevent incidents. From the employer's perspective, monitoring promotes productivity and affords better control over counterproductive employees; it is justified because employers "own" or pay for employee time and resources such as computer equipment and network connections. However, privacy rights advocates seek to ensure that employees will not suffer unwanted intrusions and that potentially harmful information will not be acquired about them. Critics note that monitoring can increase employee stress, reduce commitment, and lower productivity (Brown, 1996). Monitoring perceived as invasive with an implied lack of trust may contribute to employee job dissatisfaction, and management intervention on suspected employee disgruntlement issues may actually increase an employee's frustration level (Shaw & Fischer). Moreover, predictions imply the potential for false accusations, which can affect the career of the accused. Complicating the situation further, it has been observed that inadequate attention and action by an employer can *increase* insider activity. Such influences on trust have been described as the "trust trap" (e.g., Band *et al.*, 2006).

Employment is founded upon trust, which depends on the status of privacy, individual rights, rights of the organization, and the organization's power. Even though the organization typically asserts and society acknowledges its right to conduct electronic workplace monitoring, there is the potential for reduced trust. But if the process is disclosed fully, explained, and managed equitably across employees, it may not be considered unfair by employees, and the mutual trust relationship required for a healthy organization may remain intact. Thus, the data monitoring needed to inform a predictive psychosocial model should be done openly, with proper privacy safeguards, and based on actual behavior and events that are identified as part of the normal performance assessment process.

If behavioral data are to be monitored, what type of data should be acquired?

Personal Information. Generally, use of personal information within federal institutions is *not* likely to be appropriate or legal, no matter how useful it might be in insider threat mitigation. The employee's legal right to, and expectation of, privacy of medical records and life events such as birth, adoption, or divorce trumps the organization's desire to predict insider threat. An employee's marital and financial problems likely could not be used in a typical system. However, such life events are known to increase stress in many individuals; signs of trouble may arise not only from such personal events as divorce or death in family, but also from work-related stress due to performance issues (Band *et al*., 2006).

Manager's assessment of employee morale. An attentive manager should be mindful of an employee's personal situation and whether their behavior reflects stress or other issues. Such attentiveness provides a supportive working environment that leads to higher employee satisfaction and less likelihood of disengagement, stress, and resulting insider threat. Therefore, regardless of the personal life events that may underlie behavior, an attentive manager can provide judgments useful in a monitoring/analysis program. Further, an auditable trail of such information lets employees examine and correct any biased opinions and protects the organization from liability.

Social/Organizational Information. Unlike personal information, most work-related employee data may be used legally in observing, reporting, and correcting inappropriate or suspicious behavior. Many employees receive annual performance evaluations that may address issues about productivity, attitude, and interpersonal skills. Recurring "rule conflicts" or "personality problems" may be observed before actual insider threat events. These observations might be elements of insider threat mitigation strategies. Many suggest that managers should keep detailed records and note trends regarding events that result in employee disciplinary action (Band *et al*., 2006). Feedback obtained from "360 degree evaluations" by associates and direct reports as well as managers should be useful in assessing psychosocial factors (particularly if a manager is reluctant to provide negative feedback). Also, employee records may contain complaints by or against the employee and information related to employment applications such as education and work history.

Considering the devastating affect of a false accusation on an employee, what are the implications of the predictive approach?

A benefit of a predictive approach is the potential for an attentive manager to speak with stressed employees and possibly avert a crime by addressing underlying problems. A risk is the potential damage that may arise from false accusations. In adopting a predictive approach, there is a distinction between detection of indicators that precede a crime and detection of criminal evidence. In a predictive model, detection involves identifying precursors, not identification of the actual exploit. Indicators may be misleading due to the uncertainties of behavioral measures. Therefore, it is critically important to keep the human in the loop; a predictive system should be a tool for "tapping analysts on the shoulder" to suggest possible "persons of interest" on whom to focus limited resources. The system concept is to preserve the analyst's key decision making authority and responsibility, while helping to reduce the information load, the risk, and the cost of "false alarms."

Research on Psychology of Insiders

Research characterizing psychological profiles of malicious insiders is largely focused on case studies and interviews of individuals convicted of espionage or sabotage (Gelles, 2005; Krofcheck & Gelles, 2005; Parker, 2005; Project SLAMMER, 1990). Band *et al.* (2006) and Moore *et al.* (2008) summarize findings that reveal behaviors, motivations, and personality disorders associated with insider crimes such as antisocial or narcissistic personality. Anecdotal research is *post-hoc* (mostly derived from interviews with convicted criminals) and speculative in its predictive value. Also, assessing such personality disorders and motivations in an organization is difficult at best, and there is no guarantee that management or Human Resources (HR) staff will be able to do so accurately and consistently since a typical organization does not administer psychological or personality inventory tests. Another challenge is the fact that no studies assess and compare the prevalence of these "insider threat" predispositions with occurrence rates in the overall employee population—an important comparison needed to validate the hypothesized relationship.

Nevertheless, the body of research using case studies warrants continued efforts to address psychosocial factors. One approach is to develop predictive models that correlate the psychological profiles or behaviors that have been observed in case studies to insider crime—e.g., personal predispositions that relate "… to maladaptive reactions to stress, financial and personal needs leading to personal conflicts and rule violations, chronic disgruntlement, strong reactions to organizational sanctions, concealment of rule violations, and a propensity for escalation during work-related conflicts" (Band *et al.*, 2006, p. 15 and Appendix G). While the factors described in the extant research reflect psychological profiles inferred from case studies and interviews by staff psychologists, the present approach attempts to synthesize a set of indicators from this research and operationalize a set of corresponding observables (proxies for the indicators) that may be extracted from a manager's evaluations of staff behavior and performance. A complementary approach is to develop instructional methods to raise managers' awareness and expertise to detect the warning signs of potential insider threats: For example, US CERT (Moore *et al.*, 2008) offers workshops and interactive training; and a new R&D program at the Office of the Secretary of Defense is developing game-based methodologies to accelerate managers' awareness and ability to recognize behavioral signs of insider threat.

Additional public debate and legal guidance are needed to establish a formal policy for insider threat monitoring and mitigation. For researchers, however, it is legitimate to consider the utility of such data for insider threat mitigation, while the research community also addresses the possible impacts of these monitoring programs on the work environment. A modeling strategy that integrates psychosocial and cyber data to detect precursors to malicious exploits is described in the next section.

Predictive Modeling Approach

This section focuses on a possible structure of a predictive model combining psychosocial and traditional cyber data. Defining possible precursors to insider threat exploits in terms of observable cyber and psychosocial indicators and integrating these indicators in an analytic model is a major challenge. Confidence levels in the predictive power of psychosocial and cyber indicators can vary widely depending on managers' ability to correctly assess employees and the abilities of malicious insiders to hide their actions. Thus, the predictive modeling approach is to raise early flags to involve human analysts in the process.

A fundamental assumption is that not all possible data can be collected continuously, and some (e.g., HR records) may not be available in real time. Some data may indicate that additional scrutiny is needed but might not be linked sufficiently to insider threat to warrant specific action (for example, reprimands by management). We therefore adopt an incremental approach to data collection, analysis, and decision making in which different data are collected and analyzed for different individuals depending upon their position and insider threat risk determined by the model. Some observations or derived indicators may require immediate response if they indicate malicious actions. However, the most sophisticated insiders likely operate more subtly, hiding behaviors within "background noise" to elude detection. The model should track small changes in behavior over time to reveal trends that are discernable above the background activity.

Conceptual Design

At the highest level, the model comprises a knowledge base of indicators and heuristic models of insider behavior. Indicators are essentially the semantics of insider behavior and characteristics—interpretations of intentions and actions based on observations. This knowledge base informs all of the components of the insider threat model, and is in turn updated or modified by outputs from components that perform functions such as data collection, data fusion, and analysis. The process can be thought of as a multi-layered analysis/inference process that progresses from *Data* to *Observations* to *Indicators* to *Behaviors*, as depicted in Figure 1.

Observations are processed from cyber and psychosocial data to infer *indicators*—e.g., "excessive attempts to access a privileged database" or "presence of automated scripts." On the cyber side, observations may include registry entries, IDS/IPS events, and firewall logs. On the psychosocial side, one may infer an indicator such as "anger management" based on observations such as entries in an HR database relating to arguments with supervisors. Employees may exhibit indicators to varying degrees: someone who has difficulty recalling a recently changed password might appear on a security log as making excessive attempts to access a protected database, but someone running password-cracking software exhibits the indicator to a higher degree.

Indicators are processed to infer *behaviors*. Behaviors are sequences of activities for achieving some specific purpose, whether malicious or benign; the objective is to warn analysts about inferred behaviors consistent with established patterns of insider exploits. Example patterns include multiple policy violations indicating attempts to run unauthorized computer programs, particularly if occurring after normal working hours, and substantial downloading of files followed by emailing them. Normal work activities occasionally resemble malicious activity complicating the situation. It is thus the aggregation of these activities that needs to be recognized as potentially raising the insider threat risk. For example, as in the scenario, isolated psychosocial indicators would not point to espionage by themselves, but when issues like anger, stress, and disgruntlement are observed along with trust/risk factors such as the employee's access to sensitive information, that pattern increases risk.

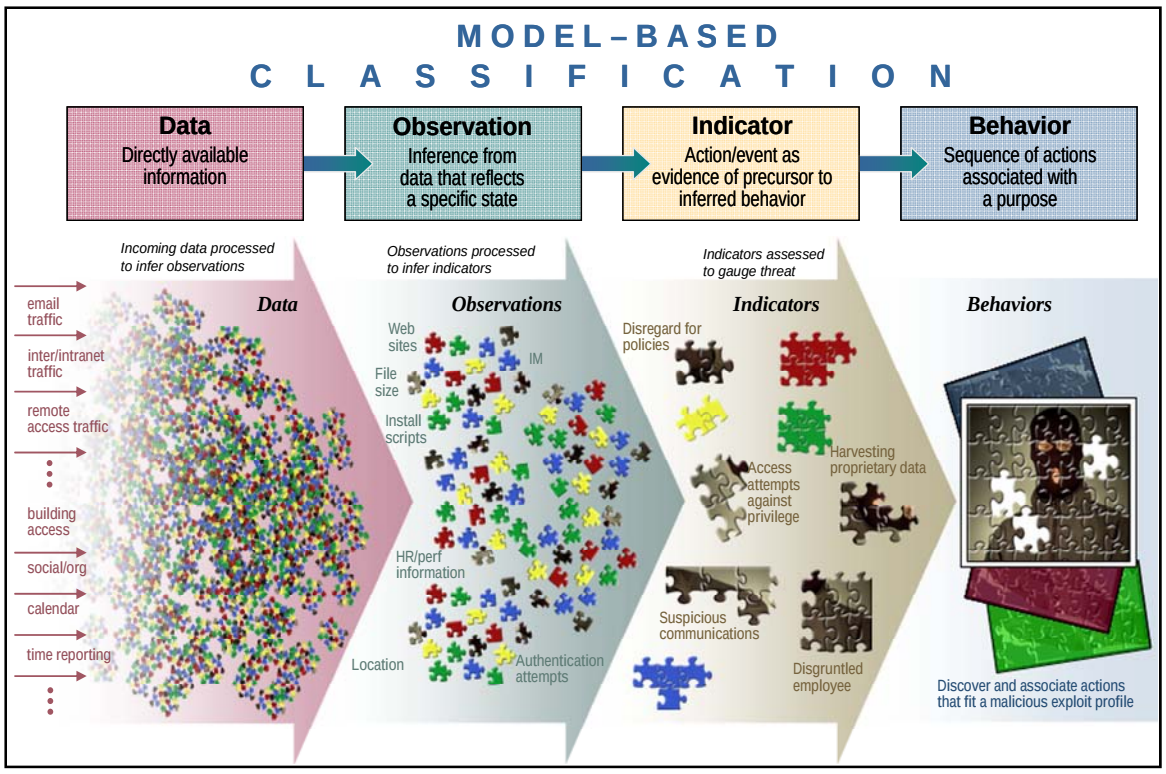

Figure 1. Model-Based Predictive Classification Concept: Incoming data processed to infer observations; observations processed to infer indicators; indicators assessed to gauge threat.

The conceptual model employs a hybrid approach based on pattern recognition and model-based reasoning. While identifying deviations from "normal" behavior—anomalies—is part of the threat analysis, so too is reasoning about conformance with prototypical behaviors that reflect possible malicious exploits. The challenge is to conduct model-based reasoning on the recognized patterns at a semantic level rather than applying template recognition.

Reasoner

A large amount of noisy data are analyzed in the transitions from *data* to *observations* to *indicators* to *behaviors*. The data are noisy because most of the tracked events are difficult to distinguish from daily work activities. The Reasoner must discern variations from norms in these events that suggest malicious actions are planned or underway.

The Reasoner processes data to infer cognitively meaningful states, or observations. The degrees to which the observations are supported are recorded as *virtual evidence (Pearl, 1988, p. 44-46,)* on nodes in a Dynamic Bayesian Network (DBN). From these, the DBN calculates belief levels assigned to *indicators*. In many scenarios, the order of events matters, as does the elapsed time between events. These temporal properties are captured by finite state machines modeled by the DBN. The Reasoner next assesses current indicators in combination with previously inferred indicators and behaviors to determine the likelihood of behaviors that represent threats. The probabilities used by the Reasoner are determined from the analyst's knowledge of threat scenarios. The model is verified by comparing the output of the network to the judgments of the analysts describing the scenarios.

Figure 2 illustrates the process. At top left is a conceptual representation of the momentary status of the employee list. A three-level status hierarchy is assumed in which baseline or nominal status applies initially to all employees, after which monitoring and analysis as described above may elevate the status of an employee to a higher risk level. In the figure, Employee222 represents a case that is similar to our example scenario.

The finite state machine is implemented within the DBN by having states and transitions represented by nodes. The transition from state S_i to state S_j via transition T_k is modeled by calculating the current belief for S_j from the previous belief for state S_i and the current belief in transition T_k. Figure 2 illustrates how the previous belief level for *Staff Activity* is used to influence the current belief in *Level II* activity through the placeholder node *Previous Staff Activity*—the dashed line indicates the temporal link. The Reasoner has the ability to *age,* or partially discount, evidence provided by previous time steps.

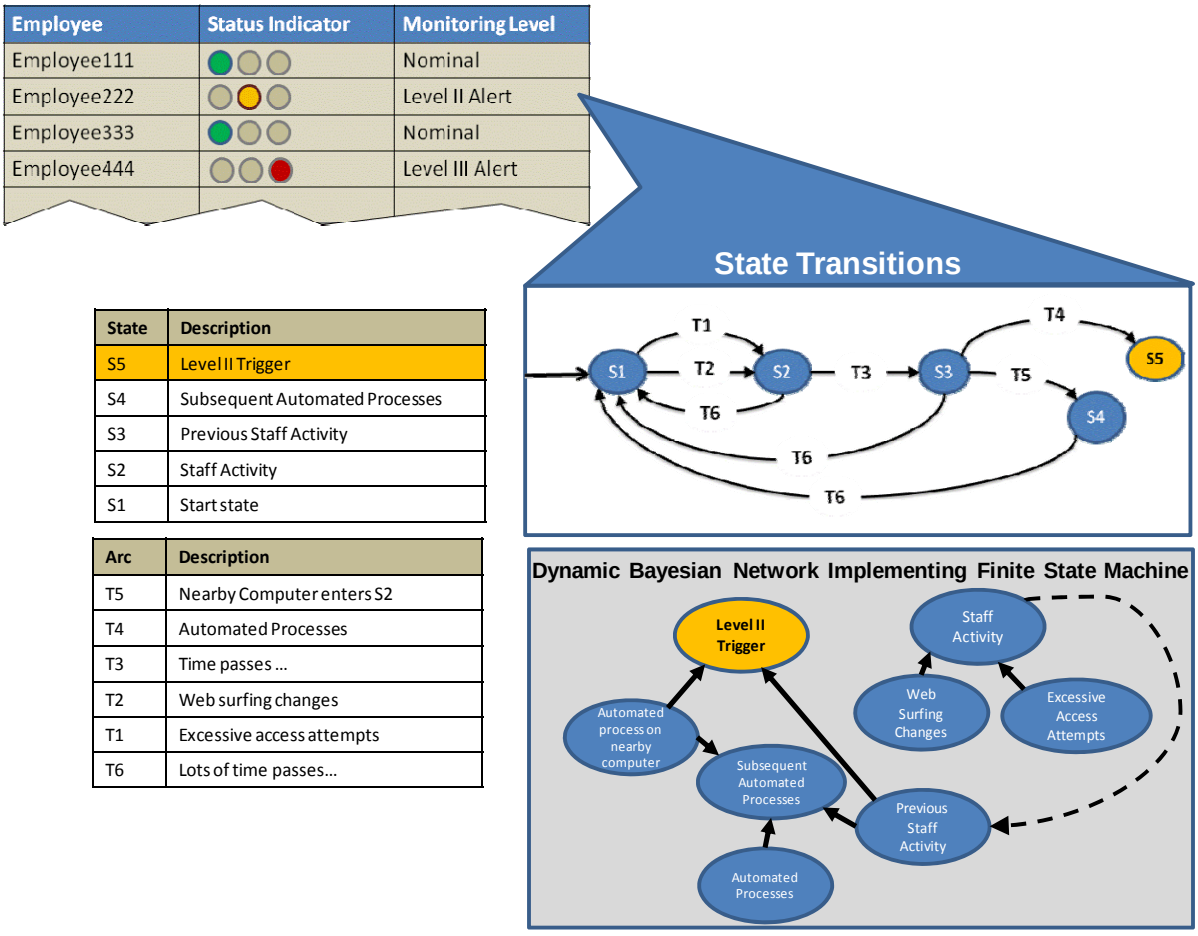

Figure 2. Illustration of the Reasoner output and its representation as a finite state machine for the portion of the model relevant to the example scenario, showing states (S), transitions (T) among states, and implementation of the finite state machine as a dynamic Bayesian network.

Psychosocial Model

The psychosocial model outputs indicators as inputs to the Reasoner. It uses a data-driven approach based on personnel data that are likely to be available. These indicators (see Table 1) are implemented as nodes in a Bayesian network. The model assigns different levels of psychosocial risk from individual

indicators and their combinations. *Disgruntlement*, *accepting feedback*, *anger management*, *disengagement*, and *disregard for authority* have higher weights and are implicated more strongly as indicators, while the indicators *personal issues, self-centered, dependability*, and *absenteeism* have lower weights that reflect weaker associations with insider threat risk.

Table 1. Indicators that determine the relative "risk level" of an individual

Indicator	Description
Disgruntlement	Employee observed to be dissatisfied in current position; chronic indications of discontent, such as strong negative feelings about being passed over for a promotion or being underpaid, undervalued; may have a poor fit with current job.
Accepting Feedback	The employee is observed to have a difficult time accepting criticism, tends to take criticism personally or becomes defensive when message is delivered. Employee has been observed being unwilling to acknowledge errors; or admitting to mistakes; may attempt to cover up errors through lying or deceit.
Anger management	The employee often allows anger to get pent up inside; employee has trouble managing lingering emotional feelings of anger or rage. Holds strong grudges.
Disengagement	The employee keeps to self, is detached, withdrawn and tends not to interact with individuals or groups.
Disregards authority	The employee disregards rules, authority or policies. Employee feels above the rules or that they only apply to others.
Performance	The staff member has received a corrective action (below expectation performance review, verbal warning, written reprimand, suspension, termination) based on poor performance.
Stress	The employee appears to be under physical, mental, or emotional strain or tension that he/she has difficulty handling
Confrontational	Employee exhibits argumentative or aggressive behavior or is involved in bullying or intimidation
Personal issues	Staff member has difficulty keeping personal issues separate from work and these issues interfere with work
Self-centered	The staff member disregards needs or wishes of others, concerned primarily with own interests and welfare.
Dependability	Employee is unable to keep commitments /promises; unworthy of trust.
Absenteeism	Staff member has received a disciplinary action (verbal warning, written reprimand, suspension, termination) for excessive time away from work

Assessment Challenges: Validating the Model

How should the effectiveness of an automated insider threat tool be assessed? No standard metrics or methods exist for measuring success in reducing the insider threat—this "capability gap" is one reason why the insider threat problem was listed second in the 2005 INFOSEC Hard Problems List (http://www.infosec-research-org/docs_public/20051130-IRC-HPL-FINAL.pdf). Other challenges are the lack of appropriate data and "ground truth" for predictive assessment, the large degree of overlap between observable behaviors of normal versus malicious activities, and the difficulty of finding population base rates.

The most rigorous form of evaluation of a predictive model is to test the predictions against a set of real cases, but due to the nature of the problem, applicable cases are rare. Further difficulties arise from the fact that data are collected over long time spans, making it difficult for experts to comprehend and reason about large volumes of data. Experts also may vary in their assessments of risk for a given set of indicators, depending on their background and experiences. In addition, while it is reasonable for experts to validate the findings of the system to perceived matches to insider threats, it is not practical for experts to examine all the observables for monitored subjects to determine which of them should be flagged. A confounding problem is that experts could find evidence of a threat that is not modeled by the system, causing difficulties in the interpretation of test results. Finally, in integrating psychosocial indicators with cyber-indicators, the model requires experts from disciplines typically outside of the experience and comfort zone of cybersecurity and counterintelligence analysts.

While an empirical test is the ultimate aim, other evaluation approaches can be used to test aspects of the model. An objective in validating the psychosocial component of the model was to demonstrate agreement between the model and expert judgments. This requires the following steps:

- Obtain expert judgments on what constitutes a valid threat, what constitutes valid indicators for that threat, and how to tie indicators to observables.
- Develop test scenarios with experts' help—scenarios must be specified in detail with appropriate data and observables that will drive the model
- Obtain expert judgments on the scenarios that will be used to test the model
- Operate the model on the data or observables associated with a scenario. The model must characterize the extent to which the observables match a scenario. These outputs are compared to experts' assessments of the same sets of observables.

Verification has been accomplished by soliciting judgments from expert evaluators who examined the same observables used by the model. Developers conducted this verification both as unit tests and as a quality control on completed models. Evaluations used case studies, similar to but more detailed than the example scenario, and additional datasets (some of which were fabricated to test specific aspects of the model, and some generated by simulation software).

Verification/Validation of Psychosocial Model. The Bayesian network model was developed from two HR experts' judgments—verified by examining the model's agreement with their judgments. Figure 3(a) shows the results of a verification test comparing the output of the psychosocial model with the (combined) judgments of our HR experts (risk judgments were provided on a 0-10 scale and then normalized to a 0-1 scale). The high value of R^2 = 0.94 indicates a good fit.

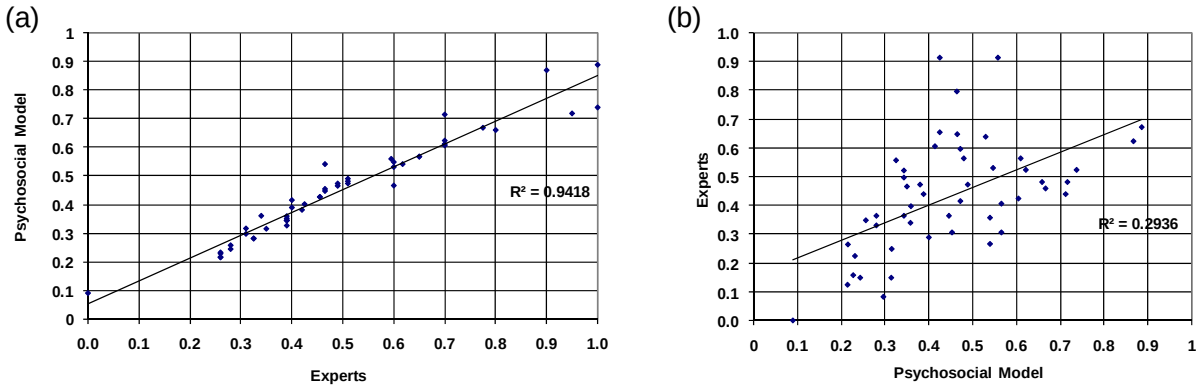

Figure 3. (a) Verification test showing the fit of the Psychosocial model to judgments of two experts used during development of the model (R^2 = *0.94*); (b) validation test showing the fit of the model to combined judgments of three different experts (R^2 = *0.29*).

The model was next validated against the judgments of three additional HR experts who were asked to judge the level of risk in 50 case studies involving different combinations of observed indicators. (This was a subset of the cases used to develop the model.) Figure 3(b) shows a scatterplot comparing the mean expert judgments with the psychosocial model's output. The resulting correlation was substantially less (R^2 = 0.29), although still significant. Much of the disagreement was traced to four cases in which five of the lesser indicators were present without any of the more serious indicators—these were judged differently by our "development" experts compared to the evaluation experts. Further investigation also suggests that the absence of more serious indicators may have been interpreted differently by the two groups of experts. Removing the four cases that did not exhibit any of the most serious indicators, the agreement *between* pairs of *expert evaluators* (ranging from R^2 = 0.33 to 0.54) was comparable to the agreement between the model's output and the individual evaluators (ranging from R^2 = 0.21 to 0.43). Additional work and more careful instructions to expert evaluators about interpreting the indicators will be necessary to calibrate the parameters of the model. At this stage, we may observe that the real-world diversity of expert evaluators can be reflected in the model. It will be valuable to preserve this diversity and allow different analysts to supply different weights or parameters when working with an operational system. This diversity also underlines evaluation challenges noted earlier.

Injection Testing (Verification) of Cyber Component. Because of the complexity of models and variability of evaluators' responses, we employed fault injection (Barbosa *et al*., 2007). The focus was to determine if the number of detections (defined as an individual whose behavior merits additional monitoring) increased with the number of simulated "bad-guys" (the injection rate).

For the injection study, data were generated for 100 simulated employees over 4 months. Simulation runs were developed for three conditions: 0, 6, and 12 simulated employees were stochastically defined as malicious insiders who would (with elevated probabilities for specific actions) engage in an exfiltration scenario involving an attempt to access and steal privileged information. The simulations were developed by staff who were not involved in implementing the Reasoner. Four simulation runs were conducted at each of the three injection rates—thus the model analyzed a total of 400 simulated employees at each injection rate. Each of the 100-employee simulation sets comprised 2.8-3.3 million

records that reflected 39 different kinds of "events" with the present analysis focusing on 10 of these event types for the specific exploit.

For each simulated employee, the model produced output ranging from lowest (0.0) to highest possible (1.0) threat value. We then assessed the association between the observed frequencies of outputs in five threat intervals (0.0-0.2, 0.2-0.4, 0.4-0.6, 0.6-0.8, and 0.8-1.0) at the three injection rates. As would be expected, most of the simulated cases analyzed by the model yielded negligible threat values (roughly two-thirds), with the remainder (406 cases out of 1200) leading to non-zero threat values. Our evaluation focused on these cases to "tease out" the model's outputs for the more challenging data. For these data, a Contingency Coefficient analysis (Siegel, 1956) uses a chi-squared significance test with 8 *df*, i.e., (5 rows - 1) x (3 columns - 1) = 8, yielding χ^2 = 575.3 ($p<<.001$). The degree of association between the two variables is obtained from the equation,

$$C = \sqrt{\chi^2/(N+\chi^2)},$$

where N = 406. The obtained value of C = 0.76 indicates a strong association. An employee was deemed "suspicious" if the threat value exceeded a threshold.

Figure 4 plots the cumulative frequencies of detections of staff, above a given threshold (threat value interval), for each of the threat value intervals at each of the three injection rates (data are transformed to frequencies per 100 to make the relationships easier to interpret for the three injection rates). The figure shows that at the injection rate of 12, threat values above 0.8 were found for five staff; 20 staff had threat values above 0.6. If we consider 0.7 as a first approximation of a threshold for "flagging" staff for further analysis, we estimate that eight staff would be "flagged" for the injection rate of "0," six staff would be "flagged" for the injection rate of "6," and 12 staff would be "flagged" for the injection rate of

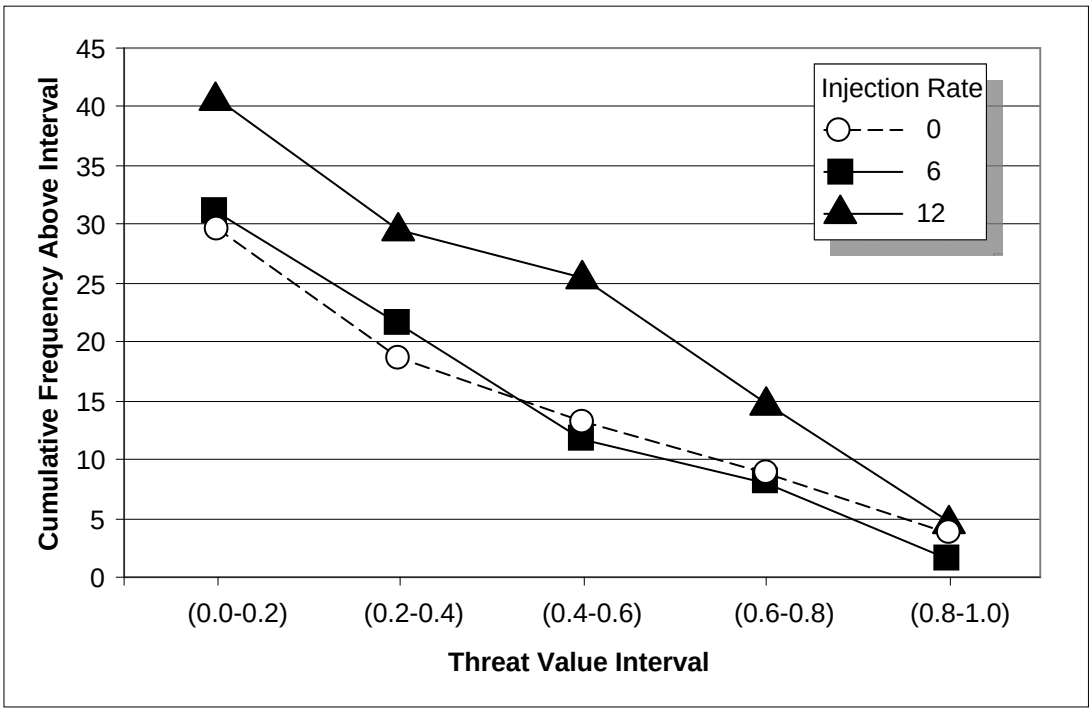

Figure 4. Cumulative frequencies above a given threat value interval as a function of injection rate.

"12." This clearly shows a trend that is consistent with the injection rate. However, the outcome for the zero injection rate condition is problematic, indicating some false alarms. It should be noted that since the flagging represents only a recommendation to increase scrutiny, this is not the same as the traditional false-positive rate associated with detection strategies. Further, the recommended additional scrutiny could reduce the threat value for simulated staff in the zero injection rate condition, thereby reducing false alarms. More challenging evaluations would assess the impact on the operational environment by tracking performance (e.g., comparing system alerts with expert judgment) and morale.

Conclusion

The insider threat—especially espionage and sabotage involving computer networks—is among the most pressing cybersecurity challenges that threaten government and industry information infrastructures. Unfortunately, data sensitivity makes progress in this area difficult. No single intrusion detection or threat assessment technique gives a complete picture of the insider threat problem. A predictive modeling approach to insider threat mitigation was described by incorporating both cyber and psychosocial data within an anticipatory decision framework. The model automates the detection of high-risk activities on which to focus and inform the analysis conducted by responsible cybersecurity analysts.

Current practice tends to be reactive as it focuses on detecting malicious acts after they occur with the aim of identifying and disciplining the perpetrator. The objective of this research is to develop a framework for a model-based system that uses psychosocial indicators as well as cyber indicators of potential abuse of network resources to predict possible malicious exploits. Some indicators may be observed directly, while others are inferred or derived from observed data. Defining possible precursors in terms of behavioral observable cyber and psychosocial indicators is a major challenge in developing a predictive methodology.

An informed and enlightened organization requires that management and HR staff be equipped with tools to maintain awareness of worker satisfaction and well-being—but not overstepping ethical and privacy boundaries—that enables thoughtful, proactive responses to situations that increase the risk of insider threat activity. Further research and technology development, as well as discussion of social and ethical issues in employee monitoring should remain among the highest priorities in addressing the insider threat.

References

Band, D. R., Cappelli, D. M., Fischer, L. F., Moore, A. P., Shaw, E. D., & Trzeciak, R. F. (2006). *Comparing insider IT sabotage and Espionage: A Model-based analysis*. Carnegie-Mellon Software Engineering Institute Technical Report CMU/SEI-2006-TR-026.

Barbosa, R., Silva, N, Duraes, J., & Madeira, H. (2007). Verification and Validation of (Real Time) COTS Products using Fault Injection Techniques. In *Proceedings of the Sixth International IEEE Conference on Commercial-off-the-Shelf (COTS)-Based Software Systems*, Feb. 26 2007-March 2 2007. 233-242.

Brown,W. S. (1996). Technology, workplace privacy, and personhood. *Journal of Business Ethics, 15,* 1237–1248.

Gabrielson, B., K. M. Goertzel, B. Hoenicke, D. Kleiner, & T. Winograd (2008). *The Insider Threat to Information Systems: A State-of-the-Art Report.* Herndon, VA: Information Assurance Technology Analysis Center (IATAC), December 2008.

Gelles, M. (2005) Exploring the Mind of the Spy. In Online *Employees' Guide to Security Responsibilities: Treason 101*. Texas A&M University Research Foundation. http://www.dss.mil/search-dir/training/csg/security/Treason/Mind.htm

Greitzer, F.L. & Endicott-Popovsky, B. (2008). Security and privacy in an expanding cyber world. Panel Session, 24th Annual Computer Security Applications Conference (ACSAC), Anaheim, CA. Dec 11, 2008.

Keeney, M., E. Kowalski, D. Cappelli, A. Moore, T. Shimeall, & S. Rogers. (2005) *Insider Threat Study: Computer System Sabotage in Critical Infrastructure Sectors.* U.S. Secret Service and CERT Coordination Center, Carnegie Mellon Software Engineering Institute, May 2005.

Kramer, L.A., R.J. Heuer, Jr., & K.S. Crawford. (2005) *Technological, Social, and Economic Trends That Are Increasing U.S. Vulnerability to Insider Espionage.* Technical Report 05-10. PERSEREC. May 2005.

Krofcheck, J.L., & Gelles, M.G. (2005). *Behavioral Consultation in Personnel Security: Training and Reference Manual for Personnel Security Professionals*.

Lane, F. S. III. (2006). The Naked Employee: How Technology is Compromising Workplace Privacy. *American Management Association*, 2006, p. 261.

Moore, A. P., D. M. Cappelli, & R. F. Trzeciak (2008). *The "Big Picture" of Insider IT Sabotage Across U.S. Critical Infrastructures.* Software Engineering Institute, Carnegie Mellon University, May 2008.

Parker, D.B. (1998). *Fighting Computer Crime: A New Framework for Protecting Information*. New York: John Wiley & Sons, Inc.

Pearl, J. (1988). *Probabilistic Reasoning in Intelligent Systems*. San Francisco: Morgan Kaufmann, 44-46.

Project SLAMMER Interim Report (U). Director of Central Intelligence/Intelligence Community Staff Memorandum ICS 0858-90, April 12, 1990. Project Slammer is a CIA-sponsored study of Americans convicted of espionage against the United States. A declassified interim report is available at: http://antipolygraph.org/documents/slammer-12-04-1990.shtml and http://antipolygraph.org/documents/slammer-12-04-1990.pdf.

Shaw, E.D. & L.F. Fischer. (2005) *Ten Tales of Betrayal: The Threat to Corporate Infrastructures by Information Technology Insiders. Report 1—Overview and General Observations*. Technical Report 05-04, April 2005. Monterey, CA: Defense Personnel Security Research Center (PERSEREC).

Siegel, S. (1956). *Nonparametric statistics for the behavioral sciences.* New York: McGraw-Hill.

Tabak, F. & W. P. Smith (2005). Privacy and electronic monitoring in the workplace: A model of managerial cognition and relational trust development. *Employee Responsibilities and Rights Journal, 17 (3)*, Sept 2005.