Identifying At-Risk Employees: Modeling Psychosocial Precursors of Potential Insider Threats

Conference Paper · January 2012

DOI: 10.1109/HICSS.2012.309

CITATIONS
78

READS
1,639

5 authors, including:

Frank L Greitzer
PsyberAnalytix
115 PUBLICATIONS **2,007** CITATIONS
SEE PROFILE

Christine Noonan
Pacific Northwest National Laboratory
28 PUBLICATIONS **273** CITATIONS
SEE PROFILE

Angela Dalton
Pacific Northwest National Laboratory
14 PUBLICATIONS **283** CITATIONS
SEE PROFILE

Ryan Hohimer
Semantic Arts Inc.
15 PUBLICATIONS **303** CITATIONS
SEE PROFILE

Identifying At-risk Employees: Modeling Psychosocial Precursors of Potential Insider Threats

Frank L. Greitzer
Pacific NW National Laboratory
Richland, WA 99352
frank.greitzer@pnnl.gov

Lars J. Kangas
Pacific NW National Laboratory
Richland, WA 99352
lars.kangas@pnnl.gov

Christine F. Noonan
Pacific NW National Laboratory
Richland, WA 99352
christine.noonan@pnnl.gov

Angela C. Dalton
Pacific NW National Laboratory
Richland, WA 99352
angela.dalton@pnnl.gov

Ryan E. Hohimer
Pacific NW National Laboratory
Richland, WA 99352
ryan.hohimer@pnnl.gov

Abstract

In many insider crimes, managers and other coworkers observed that the offenders had exhibited signs of stress, disgruntlement, or other issues, but no alarms were raised. Barriers to using such psychosocial indicators include the inability to recognize the signs and the failure to record the behaviors so that they can be assessed. A psychosocial model was developed to assess an employee's behavior associated with an increased risk of insider abuse. The model is based on case studies and research literature on factors/correlates associated with precursor behavioral manifestations of individuals committing insider crimes. To test the model's agreement with human resources and management professionals, we conducted an experiment with positive results. If implemented in an operational setting, the model would be part of a set of management tools for employee assessment to identify employees who pose a greater insider threat.

1. Introduction

The insider threat refers to harmful acts that trusted individuals might carry out; for example, something that causes harm to the organization, or an unauthorized act that benefits the individual. The insider threat is manifested when human behaviors depart from established policies, regardless of whether it results from malice or disregard for security policies. The most serious crimes and abuses include espionage, sabotage, terrorism, embezzlement, extortion, bribery, and corruption. Malicious activities also include copyright violations, negligent use of classified data, fraud, unauthorized access to sensitive information, and illicit communications with unauthorized recipients. Insider criminals do not all have the same motivations and characteristics; our main focus is on those associated with espionage, IT sabotage, and theft of intellectual property.

The annual e-Crime Watch Survey conducted by Carnegie-Mellon's CERT program reveals that for both the government and commercial sectors, current or former employees and contractors pose the second greatest cybersecurity threat, exceeded only by hackers; the financial impact and operating losses due to insider intrusions are increasing [1, 2].

Among the most significant challenges facing the development of a predictive analytic methodology for detection, prediction and mitigation of insider threats is (1) defining precursors—events prior to the attack)—in terms of observable cyber and psychosocial indicators; and (2) developing a methodology that integrates these indicators [3]. The focus of the present paper is on this challenge.

2. Relevant Research

There has been much research into the psychology and motivation of insiders, but the fact remains that it is difficult to predict who will commit security fraud [4]. Shaw and Fischer [5] reported that nine of 10 cases studied involved serious employment crises and that in nearly every case the subject of the study exhibited signs of disgruntlement and serious personnel problems months prior to an attack. These subjects reacted to off-line personal conflicts, stresses, and disappointments through electronic behavior. One of the most important findings of this research was that there was a window of opportunity for dealing with the personnel problems affecting these subjects. These individuals were reportedly disgruntled in some cases for over a year prior to their attacks, and management was aware of these personnel problems weeks, if not months, prior to the attack. Thus Shaw and Fischer concluded that most of the threats in their study could have been prevented by timely and effective action to address the anger, pain, anxiety, or psychological impairment of perpetrators. Yet there were consistent intervention problems. In fact, in many cases ill-considered management actions exacerbated the

978-0-7695-4525-7 2012

DOI 10.1109/HICSS.2012.309

problem. This finding indicates the need for improved management training and procedures covering interventions with at-risk individuals [5].

Despite these compelling observations, in current practice no systematic methods are available to evaluate psychosocial behaviors that can predict increased risk for insider threats. The present work follows recommendations by Schultz [6] to develop a new framework for insider threat detection based on multiple indicators that not only address workstation and network activity logs but also include preparatory behavior, verbal behavior and "personality traits."

Varied reasons why an employee would turn against an employer include a desire for revenge, disgruntlement about one's job, and financial problems, to name a few. A desire for revenge may be driven by the satisfaction of causing costly damage to the corporation but it can be also include a motive of financial gain (e.g., by exploiting intellectual property). In either case, the employee may have exhibited stress or some form of dissatisfaction about his or her circumstances. These factors, if properly evaluated in a timely manner, could alert an organization about a developing insider crime. Identifying employees who show elevated risk of insider threat can also serve to provide help the employee before a bad situation turns worse. Thus, a psychosocial model benefits both the employees and the employers if incorporated as a tool in regular staff evaluations, and if the predictions of the tool are acted on appropriately.

3. Psychosocial Model Description

Implementation of the psychosocial reasoning used a data-driven approach based on personnel data that are likely to be available (a discussion may be found in [7]). The indicators used in the model, such as disgruntlement, anger management issues, and disregard for authority, are defined in Table 1. As discussed in [7], the selection of these indicators reflects an approach that (a) acknowledges privacy considerations that limit access to private information that has been associated with insider crime (such as financial and medical records) and (b) relies on observable behaviors rather than psychological personality predispositions that would otherwise have to be determined through personnel evaluations. We developed the list of indicators based on examination of published case studies and discussions with experienced human resources (HR) professionals.

The indicators in Table 1 contribute differentially to the judged level of psychosocial risk—

Table 1. Psychosocial Indicators

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.
Not 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 Issues	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; avoids meetings.
Disregard for Authority	The employee disregards rules, authority or policies. Employee feels above the rules or that they only apply to others.
Performance	The employee 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 Behavior	Employee exhibits argumentative or aggressive behavior or is involved in bullying or intimidation.
Personal Issues	Employee has difficulty keeping personal issues separate from work, and these issues interfere with work.
Self-Centeredness	The employee disregards needs or wishes of others, concerned primarily with own interests and welfare.
Lack of Dependability	Employee is unable to keep commitments /promises; unworthy of trust.
Absenteeism	Employee has exhibited chronic unexplained absenteeism.

disgruntlement, difficulty accepting feedback, anger management issues, disengagement, and disregard for authority have higher weights than other indicators, for example. Since judgments based on observations will necessarily be subjective, there is no expectation that an objective test instrument will emerge from this research. But with appropriate training, we believe that management and HR personnel would better understand the nature of the threat and the likely precursors or threat indicators that may be usefully reported to cyber security officers.

Most importantly, the approach in predictive modeling is to provide "leads" for cyber security officers to pursue in advance of actual crimes, without which they would likely have little or no insight from which to select higher-risk "persons of interest" on which to focus analyses. For security analysis purposes, only individuals about whom a manager is "highly concerned" would be considered for further analysis in the insider threat model.

As the model's assignment of risk level increases for an individual, so too would the level of monitoring and analysis for that individual.

3.1 Bayesian Network Model

The initial Bayesian Network (BN) model was developed from judgments of two HR experts in several knowledge engineering meetings. The psychosocial indicators and the psychosocial risk were implemented as binary variable nodes in a BN model [8, 9] using Genie (GeNie 2.0.3006.0, Decision Systems Laboratory, University of Pittsburgh). The development of a BN requires several steps: First the network is constructed with linked conditionally dependent random variables that each takes on values from a domain. In our model, these values are True or False, corresponding to whether the indicators were observed severe enough to be a concern. Second, prior probabilities (priors) are assigned to each random variable. These priors, as estimated by HR experts, reflect the frequencies at which random variables take on values from their domains. For example, the prior probability that an employee is observed to exhibit severe stress is denoted P_{Stress}, and the complementary case (employee does not exhibit severe stress) is $1 - P_{Stress}$). The third step in developing a Bayesian network is to determine the influence of the random variables on the risk output to be encoded. One way to do this is ask HR experts to estimate numeric values directly in the conditional probability table (CPT) of the BN's Psychosocial Risk random variable node. Clearly this involves a large number of complicated judgments in which various numbers of factors are combined. (Formally, the total number of possible combinations is referred to as the power set P(S), the set of all possible subsets of the set S, which is 2S. In the present case, there are 4096 possible cases.)

Because this method was highly impractical, we used an alternative approach to derive the conditional probabilities through a training methodology that acquires expert judgments for only about 3% of the total number of cases and that builds on the judgments of individual priors. We constructed 110 different scenarios where employees had zero to five indicators set to TRUE, and then asked the HR experts to assign insider threat risk levels to employees who would exhibit those behaviors. The Genie expectation maximization algorithm was applied to these expert-assigned risks for the scenarios to set the conditional probabilities in the BN's CPT. This step transferred the expertise of our HR experts to the BN with the expectation that the network should predict the same insider risk in employees as the HR experts would for novel employee evaluations.

Table 2 shows the priors for observing the employee behaviors in a year as estimated from two of our HR experts. The table also shows relative judgments of the weight of each indicator in influencing one's risk assessment for insider threat when the indicator is observed alone. It is evident that (extreme) disgruntled behavior occurs relatively seldom (0.025) but has a high influence on the associated insider risk (0.400), while (extreme) self-centeredness occurred relatively often (0.100) but has a lower influence (0.180).

Table 2. Priors and weights for the model

Parameter	Prior	Weight
Disgruntled	0.025	0.400
Accepting Criticism	0.060	0.280
Anger Management	0.019	0.260
Engagement	0.040	0.310
Disregards Rules	0.075	0.340
Performance	0.020	0.160
Stress	0.030	0.200
Confrontational	0.063	0.120
PersonalIssues	0.080	0.140
SelfCentered	0.100	0.180
Dependability	0.038	0.060
Absenteeism	0.010	0.060

Intuitively, one can conclude that extreme self-centeredness, when observed alone in 10% of employees, should not cause alarm for an insider risk; otherwise an employer would have to conduct comparatively high levels of insider threat monitoring

on 10% of its workforce. (The priors differ for different labor forces and corporations—the numbers shown in Table 1 are illustrative estimates by the HR experts with experience in the domain of research scientists.)

It is interesting to note that when the priors were solicited from the HR experts, the experts were initially asked to provide the priors as probabilities. An examination of these priors and discussion with our experts suggested that these initial estimates were inflated. Recognizing that there may be certain biases associated with probability estimation (particularly for rare events with negative consequences or utilities—e.g., [10]), we therefore asked the HR experts to estimate the number of cases that occur per year in which an employee exhibits a given indicator, assuming a baseline context of about 4000 employees (consistent with their experiences at our institution). The rephrased question format appeared to have provided better estimates of the priors.

3.2 Verification Experiment

In a verification study, we asked the two HR experts used for model development to judge the severity of 61 cases on a 0-10 scale and compared their averaged scores with the output of the model (the Bayes probabilities were normalized to a 0-10 risk scale). The results, shown in Figure 1, indicate a high level of agreement between the expert judgments and the psychosocial model ($R^2 = 0.920$).

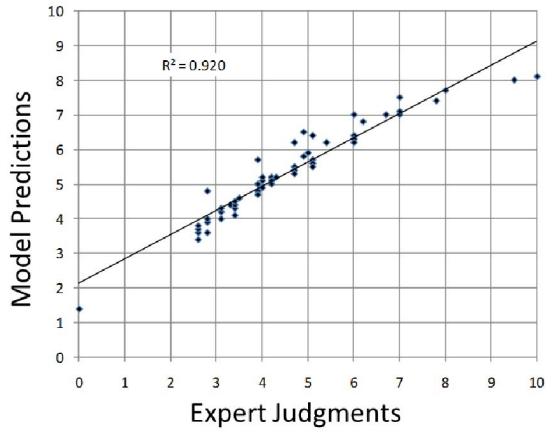

Figure 1. Verification test showing the fit of the Psychosocial model to expert judgments used to develop the model.

4. Formal Study

The analysis described in the previous section provided a degree of verification that the psychosocial model represented the judgments of experts whose judgments were used to construct the model. A more rigorous assessment was obtained by examining the degree to which the model fits judgments of another set of experts. (We recognize that the strongest form of validation would entail the ability of the model to predict actual insider exploits rather than to generate predictions that are consistent with expert judgments. A longitudinal study is required in which data are collected over a period of time and then predictions of the model are compared to actual observed events. This was beyond the scope of the present study.)

4.1 Participants and Procedure

Ten staff members were recommended by HR management for their breadth and depth of experience. Informed consent was obtained for these volunteers (IRB# 2008-05E). Each participant attended one of three 2-hour sessions conducted over a day and a half, during normal business hours. Three or four participants attended each session in which the experimenter first spent ten minutes describing the general insider threat problem and explained that the focus of this study was on behavioral factors, acknowledging that a complete response to the insider threat problem would also require integration of behavioral analyses with workstation/electronic monitoring approaches. Next, it was explained that the focus of this study was on malicious insiders, not on individuals who inadvertently propagate malicious exploits by others from the outside, such as phishing attacks and the like. After this discussion, the participants were asked to read a one-page description of the problem that included a table identifying and describing twelve behavioral factors that were studied. The experimenter explained that these factors were of interest, and the opinions of the participants were sought to help validate the importance of each of the factors. Thus, it was noted that there was no expectation that these factors contribute equally to identifying and predicting potential insider threats. With no additional discussion or questions by the participants, the experimenter described the desired procedure that the participants were asked to follow, as described below.

Each participant was given a collection of 24 cases, printed on separate pages, with the case presented both in tabular and narrative form. Figure 2 shows an example of a case that exhibited three indicators, ***Disgruntlement, Anger Management Issues, and Disregard for Authority***. Each set of 24 cases was shuffled prior to the experiment session so there was no consistent order of the cases in the sets given to the participants. The participants were asked to sort the 24 cases into up to ten categories, ranging from "no

Case 39567.

Indicator	Observed?	
Disgruntlement	Yes	
Not Accepting Feedback		No
Anger Management Issues	Yes	
Disengagement		No
Disregard for Authority	Yes	
Performance		No
Stress		No
Confrontational Behavior		No
Personal Issues		No
Self-Centeredness		No
Lack of Dependability		No
Absenteeism		No

Adam was sure he would be picked for a one-year offsite assignment that he wanted very much. Not receiving the assignment made Adam disappointed and angry at management. After Adam's manager observed that he continued to hold a grudge and exhibit the anger (**Disgruntlement, Anger Management Issues**), the group manager entered these observations into Adam's personnel folder. Following this incident, the group manager received word of Adam breaking company rules, possibly in defiance of its management (**Disregard for Authority**). No other risk indicators have been observed or recorded in his personnel folder. Other than these indicators, no additional issues have been observed or documented.

Method Used to Develop These Cases:
There are 4096 possible combinations of the 12 indicators (being observed or not). In previous model development and testing, we created a random subset of 110 cases that have from zero to 5 indicators set to "Observed." For our sorting task, we chose a representative 24 from the 110 cases and we wrote a narrative (as shown at left) for each. Narratives were vetted with one of our HR experts prior to use.

Figure 2. Example of a case used in the sorting study.

concern" on the left to "highest concern" on the right. The sorting categories were not labeled, and participants were instructed that they must use the first and last category, but they were not required to use any category in between.

Participants worked on their own and at their own pace. When the sorting task was finished, the experimenter asked the participants to further rank-order any cases that appeared in the same category. At the conclusion of this task, the participants were asked to rank-order the twelve factors, from highest to least importance, as lone indicators or predictors of potential insider threat risks.

4.2 Results

For measurement/identification purposes, we labeled the ten sorting categories using the numbers 0-9. Each case was coded into its respective sorting category (0-9), with additional discrimination using the rankings within categories, as follows: The first case in category 0 (no concern) was assigned a score = 01; the second case in that category was given score = 02; and so on. For category 2, the cases were assigned scores of 21, 22, … depending on the number of cases contained in that category. Similarly, scores were assigned for all 24 cases. For the indicator ranking task, the twelve indicators were assigned numeric/integer ranks from 1 to 12, based on the rankings assigned by the participants.

Table 3 shows the rank orders of the 12 indicators as given by the participants. The average standard deviation for the twelve indicators is 1.69, indicating a good consensus among the participants for the indicators. A few indicators stand out as having a large range among the participants: e.g., the ranks of Disengagement range from 1 to 12. As may be seen in Table 3, eight of the ten participants seemed to agree on a middling ranking (corresponding to the average rank of 5.2), while one participant considered the indicator to be the most important and another participant considered it the least important. Although the participants in the tests were given some guidelines for the definitions of the indicators, some subjective interpretation seemed to persist at the time of testing.

Variations in the participants' rankings are expected based on their different experiences and perceptions. Inspection of Table 3 shows that Self-Centeredness has the highest standard deviation (3.44). This indicator also has the highest prior (0.100) in

Table 3. Indicator rank orders by ten experts

Indicators\Expert	1	2	3	4	5	6	7	8	9	10	Average	StdDev
Disgruntlement	1	1	6	6	3	1	5	2	1	1	2.7	2.16
Accepting Feedback	6	6	8	5	5	6	6	8	6	5	6.1	1.10
Anger Management	5	3	2	3	4	2	2	4	4	3	3.2	1.03
Disengagement	4	4	5	7	1	12	7	5	3	4	5.2	2.97
Disregard for Authority	3	2	1	1	2	3	3	1	2	2	2.0	0.82
Performance	8	11	12	8	7	8	10	9	10	6	8.9	1.85
Stress	7	7	9	10	9	10	8	7	5	7	7.9	1.60
Confrontational	2	5	3	4	5	4	4	3	7	8	4.5	1.84
Personal Issues	10	10	10	11	8	9	9	12	11	9	9.9	1.20
Self-Centeredness	9	9	4	2	11	5	1	6	8	10	6.5	3.44
Lack of Dependability	11	8	7	9	10	7	12	10	9	11	9.4	1.71
Absenteeism	12	12	11	12	12	11	11	11	12	12	11.6	0.52
											Average:	1.69

Table 2. These values would likely vary depending on the type of organization and the work experience of the HR and management participants providing the estimates. We selected our participants for their many years of HR experience in diverse work environments; it may not be too surprising to find a relatively high degree of self-centeredness among high-achieving scientists.

Table 4 shows the insider threat risks assigned by the participants to our 24 test cases. The assigned risk levels range from 1 to 93, and were computed as explained above. The table shows that the average standard deviation is a low (14.62) for the 24 cases, suggesting that the ten participants were relatively consistent in judging the risk for the individual cases. The lowest standard deviation is 6.77 for case 21 and the highest is 27.33 for case 7, which appears to have been interpreted differently (case 7 was assigned risks of 1 and 2 by two participants, respectively, versus 73 and 74, respectively by two other participants.

Table 4. Rankings of the 24 test cases

Case\Expert	1	2	3	4	5	6	7	8	9	10	Average	StdDev
1	82	73	55	81	61	85	71	64	82	13	66.7	21.41
2	71	41	34	94	91	73	41	91	61	61	65.8	22.19
3	72	61	51	61	82	72	51	51	64	63	62.8	10.35
4	74	72	52	51	72	61	23	81	44	62	59.2	17.30
5	73	71	54	62	54	51	61	72	52	41	59.1	10.59
6	31	42	61	32	81	42	22	31	53	51	44.6	17.51
7	11	74	31	11	73	1	33	2	45	11	29.2	27.33
8	61	43	53	91	93	32	21	73	63	44	57.4	23.71
9	32	62	57	71	41	71	42	41	81	22	52.0	19.24
10	41	51	91	82	52	91	81	53	91	82	71.5	19.79
11	63	83	56	95	62	82	52	63	71	81	70.8	13.89
12	91	92	93	92	71	81	82	62	93	91	84.8	10.79
13	3	22	21	22	22	52	12	12	2	1	16.9	15.07
14	12	33	24	64	21	21	14	22	43	2	25.6	17.53
15	4	11	1	34	13	12	32	11	31	3	15.2	12.52
16	13	21	2	1	1	13	11	1	11	12	8.6	6.93
17	81	82	92	96	92	92	91	82	92	92	89.2	5.37
18	21	81	32	63	23	62	34	52	62	43	47.3	19.86
19	2	23	5	33	11	14	13	23	42	21	18.7	12.34
20	62	91	33	41	83	31	44	42	72	52	55.1	21.01
21	1	2	3	21	3	11	1	13	1	4	6.0	6.77
22	5	1	4	31	2	41	31	21	21	5	16.2	14.65
23	14	31	23	93	51	83	45	61	41	31	47.3	25.47
24	51	32	22	42	53	84	43	71	51	42	49.1	17.90
											Average:	14.62

The inter-rater agreement on the 24 scenarios was high (pairwise mean Pearson correlation coefficient = 0.684, standard deviation = 0.095; intra-class correlation coefficient = 0.651 where 1.0 is a perfect agreement, and inter-rater reliability coefficient with Spearman-Brown correction = 0.949; the nonparametric, Kendall's w (Kendall's coefficient of concordance) is 0.707 ($p < .001$) where 0 indicates no agreement and 1, perfect agreement). The coefficient of concordance suggests there is a high level of agreement among the raters and the agreement is statistically significant.

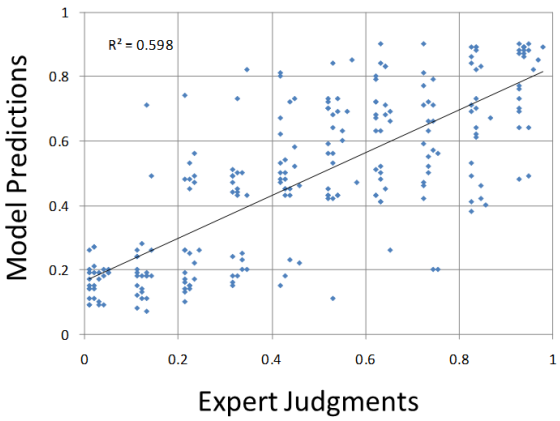

Figure 3. Bayesian prediction of 24 unique cases for a total of 240 test cases (Round-Robin testing)

4.3 Test of Bayesian Network Model

The BN was tested using a round robin procedure, leaving out the 24 cases from one rater for testing while the 24 cases from each of the other nine raters were used in Genie's expectation maximization algorithm to learn the probabilities in the conditional probability tables in the network. Figure 3 shows the Bayesian network predictions for the 240 cases left out in testing. The scatter plot shows a clear division vertically between predictions ~0.3-0.4. Above this boundary there is at least one behavioral indicator observed from the group of the five most important indicators; and below the boundary there are only behaviors that appear less indicative of insider threat risk. Although the experts' ratings of the cases do not show this separation explicitly, the BN learned to deduce this pattern from the experts' predictions (R^2 = 0.598, RMSE = 0.188).

4.4 Other Models

Besides the BN psychosocial model, three other models were developed and tested to predict the risk a

staff member would pose if subsets of the indicators had been observed. Like the BN model, a linear regression (LR) and a feedforward artificial neural network (ANN) were trained and tested using round robin testing. In addition, a simple Counting model, was defined that counts the number of indicators observed; the Counting model did not require estimates of weights for the individual indicators and thus no training or round robin was needed.

Figure 4 shows that the counting model yielded a Pearson R^2 = 0.253 (RMSE = 0.260). The relatively poor performance is attributed to the fact that all indicators were weighted equally in this model while the experts clearly considered some indicators more important than others. This confirms the consensus in the ranking of indicators shown in Table 3—the experts considered the indicators as having different weights in predicting the risk.

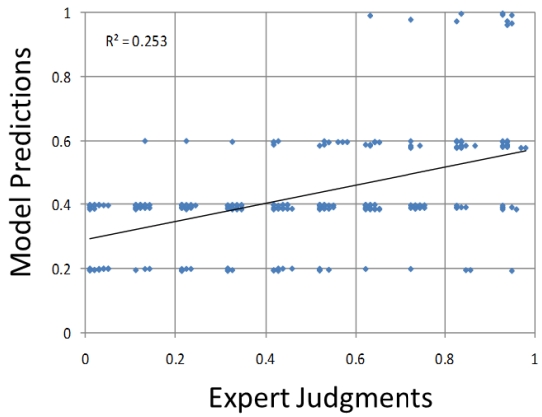

Figure 4. Counting model. The model counts unweighted behavioral indicators.

The LR model had a specific weight for each indicator. Two methods were tried when developing this model. First, a genetic algorithm [11, 12] was configured to optimize the indicator weights to produce the risk measures assigned by the experts for the cases. Second, a feedforward artificial neural network (ANN) was configured without a hidden layer (12 inputs, 1 output) and trained with the backpropagation algorithm [13-15]. The fully-trained ANN yields the weight for each indicator. As both methods optimize the weights, they produced the same weights. (Observe that round robin was used for both methods as explained above.) The performance of the LR model was R^2 = 0.592, a substantial improvement over the counting model, further confirming that the individual indicators should be weighted differently. The ANN with one hidden layer (12 inputs, 2 hidden nodes, and 1 output) was tested to discern if the problem had nonlinear properties. The performance increased only slightly to an R^2 = 0.606 (RMSE = 0.185), suggesting the problem is mostly nonlinear.[1] Table 5 summarizes the results for the individual models. Clearly the BN, the LR, and the ANN models that differentially weigh the individual behavioral indicators perform best.

Table 5. Performance of the Models

Model	R-square	RMSE
Bayesian Network	0.598	0.188
Counting Model	0.253	0.260
Linear Regression	0.592	0.191
Artificial Neural Network	0.606	0.185

4.5 Discussion

The BN scatter plot (Figure 3) shows significant gaps in how the model assigned risks (similar gap was found in the ANN scatter plot; see [16]). These gaps coincide with low and high risk indicators. When the indicators were developed, certain indicators were deemed more significant than others. The low risk indicators according to the experts in the development team were those that can be observed alone in staff members without them posing a risk. Only when these are observed together with high risk indicators, do they increase the risk. Below the gaps in the graphs, there are only combinations of the low risk indicators. Above the gap, there are one or more high risk indicators combined with the low risk indicators. None of these relationships was explicitly discussed with the test participants; the models revealed this relationship without it being observed in the scores given by the participants. One possible interpretation for the gap is that it provides a natural threshold for deciding when staff members should be monitored more closely.

Both the BN and the ANN capture some nonlinearity in the data that the LR model fails to capture. Even though the R^2 value for the BN model is a little lower than that of the ANN model and a little higher than that of the LR model, the BN has important advantages: (a) The BN is better suited to work with missing values; if behavioral indicators are neither observed/confirmed as true or false, then the BN will use the prior probabilities of those indicators to make the best predictions. A regression model, either linear or nonlinear (like the ANN), typically requires that each indicator is set to true or false for binary values. (b) The BN gives probability estimates, assuming true risk rates and priors were available during model development (the other models do not). (c) The BN provides simpler interpretations of risk values.

[1] Variations of the number of hidden nodes were tested: More than two hidden nodes did not improve performance suggesting the problem has a low complexity. Observe that one hidden node is the same as a linear model.

The modeling results show that an R^2 of about 0.6 is achievable in an expert system that simulates the consensus of ten experts. No judgment is made here as to the accuracy of the experts' opinions in predicting threat, but we observe that their consolidated judgments represent many years of experience in managing human resources. We suggest that the "average" risk predictions generated by a model representing this consolidated wisdom is better due to possible information processing limitations, individual biases, or varying experiences of individual experts. An expert system model also enables the automatic screening of staff members that is consistent and independent of the experiences held by an individual HR staff member.

5. Application

The insider threat poses a very hard detection problem and an even harder prediction problem. We have described research that suggests that any attempt to seriously address the insider threat, particularly through proactive means, must consider behavioral indicators in the workplace in addition to the more traditional workstation monitoring methods. Recognizing behavioral indicators is difficult and requires training, but we suggest that raising managers' and HR staff's awareness and skills in recognizing potential risks can only improve their effectiveness in dealing with everyday workplace challenges as well as severe insider threat risks. For a very large organization with many thousands of employees, it is difficult and costly to train sufficiently many experts so that all employees' risks are regularly and consistently analyzed. Employing such expertise in a computer-based decision aid will help ensure that the "system" is applied consistently and fairly. The model automates this process, given that the organization implements employee evaluation processes that deposit behavioral assessments in a database of personnel files at regular intervals such as during employee performance evaluations or when these behaviors are observed.

Good managers and HR staff are well aware of incidents and issues relating to "concerning behaviors"–e.g., increasing complaints to supervisors regarding salary, increased cell phone use at the office, refusal to work with new supervisors, increased outbursts directed at coworkers, and isolation from coworkers [17]—and for the most serious occurrences, which are the focus of our model, there will be communications/discussions among HR staff and management. Management is not only aware of the most egregious behaviors, but indicators of concerning behaviors may appear 1 to 48 months before the attack [5]. This provides a window of opportunity during which employers' awareness of risk linked to effective interventions could reduce the threat of an attack [5, 18]. Randazzo et al. [19] reported that eighty percent of insider cases in their study raised official attention for concerning behaviors such as tardiness, truancy, arguments with coworkers, and poor job performance; and in 97% of those cases, supervisors, coworkers, and subordinates were aware of these issues. However, typically there is no formal infrastructure for recording and tracking such behaviors, except when they become critical to the point where disciplinary action is taken. A system for collecting and tracking reported concerning behaviors will enable objective examination of these data and their integration with physical and cyber monitoring data to derive a complete picture of potential "problem employees" and insider threats.

It has been argued that insider threat assessment based on screening of personal characteristics will be imperfect because malicious insiders do not share a common profile and that characteristics of bad actors are shared by good actors; and "Because the set of malicious insiders is small and diverse, no single personal characteristic or set of characteristics can act as a reliable predictor of future misbehavior," [20]. We do not advocate a model based only on personal characteristics, but rather a model that integrates multiple sources of data—consistent with Schultz [6] advocacy for systems that monitor and analyze numerous clues of diverse types, including personal characteristics and suspicious cyber activities.

Because consistency and objectivity are of paramount importance in providing this type of input, managers who supply this information must be given guidelines and effective training on recognizing psychosocial indicators. A system developed to record these judgments would store data only for the most extreme cases involving grave concerns about the employee's behavior. These judgments are combined with other behavioral data that may be available (e.g., objective data such as job changes, being passed over for promotion, or disciplinary actions that may affect employee attitudes; and other possible for psychosocial triggers that may be extracted from word use in social media [21] [22]); and with cyber/workstation monitoring data to produce a composite picture that a security analyst can examine. Analysis of outputs from a psychosocial model and other more conventional workstation activity monitoring would be used in informed decisions of a multidisciplinary team comprising management, HR, security, cybersecurity personnel, as well as a counterintelligence officer for the most serious transgressions. Most importantly, the automated decision aid would be used only to inform and advise—not to invoke unilateral sanctions.

6. Current Research

We have observed that security, cyber security, and counterintelligence analysts typically use a number of forensic tools that monitor different types of data to provide alerts about suspicious activities (such as failed authentication attempts). In general the analyst has the critical and difficult data fusion or integration responsibility to recognize patterns of suspicious behaviors across disparate data domains (sensemaking tasks). Our current research focuses on the development of a reasoning system that is better suited to this sensemaking/pattern recognition process; the aim is to reduce cognitive load of analysts by inferring patterns of behavior that bridge multiple types of events and performing a higher level "triage" on alerts.

The reasoning system, called CHAMPION, for Columnar Hierarchical Auto-associative Memory Processing In Ontological Networks [23], comprises a hierarchical framework of reasoners organized by a semantic layer that enables graph-theoretic pattern recognition methods. In contrast to typical semantic graph approaches with a monolithic reasoner that is required to reason over *all* the concepts represented in the entire semantic graph, each reasoner in the CHAMPION belief propagation network reasons only about a small set of relevant concepts. Each CHAMPION reasoning component is governed by an ontology (derived from expert knowledge); the reasoning framework executes an abstraction process that successively analyzes higher-order patterns while bridging multiple data domains. This will help the analyst correlate data over space and time and across varied data sources, decrease the cognitive load on the analyst and focus attention on activities that present the most critical potential security risks.

The model can incorporate multiple data types that are generated from a variety of employee monitoring "appliances" that are currently offered in commercial products, from security event and information management system logs to web logs to data loss protection/prevention systems; the model can also incorporate data generated from behavioral and psychosocial input (although the latter data require more formal methods of collection, as discussed earlier in this paper). The data are ingested into the system's semantic graph by separate data-ingest components ("reifiers") that are tailored to the appliance/data entering the system. A conceptual illustration of a high level architecture is shown in Figure 5.

Another facet of our ongoing research aims to determine if it is feasible to extract and infer psychosocial factors from analysis of word use in social media ([21] [22], which could augment or replace subjective behavioral assessments described in the present research.

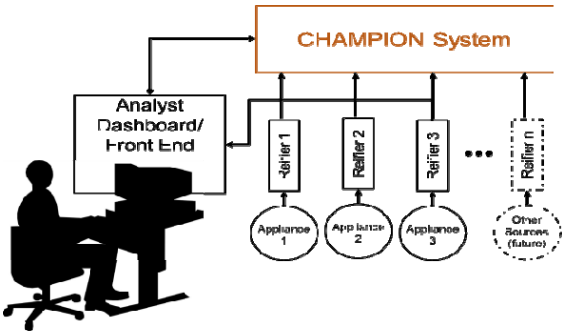

Figure 5. CHAMPION high level architecture

7. Conclusions

To protect both employees and employers, systematic methods are needed to reduce the risk of deliberate attempts to harm corporate interests or individuals. In a preponderance of reported case studies, the malicious intent of the perpetrator was "observable" prior to the exploit. We developed a prototype psychosocial model that takes this research literature into account through behavioral indicators of increased insider threat risk. We have shown that a predictive model of insider threat risk can be developed to produce predictions that are highly correlated with expert HR judgments. Four versions of a predictive psychosocial model were developed. The results showed that predictions of a Bayesian model, a nonlinear feedforward ANN model, and a linear regression model provided good fits to the judgments of ten experienced HR staff and significantly outperformed a simple counting model. Arguably, the BN model implementation might be preferred since it would be more robust against missing data and it provides easier interpretations of risk. An unequivocal conclusion is that the posited behavioral indicators have differential predictive weights.

A combination of systems that monitor user cyber data on computers/networks and a system that records psychosocial/behavioral indicators can provide a comprehensive solution to empower a HR/cyber/insider threat team with enhanced situation awareness. This will transform a reactive/forensics based approach into a proactive one that will help identify employees who are at greater risk of harming the organization or its employees. With proper privacy safeguards in place, this also provides a fair and consistent approach to employee monitoring that benefits employees and employer alike.

8. Acknowledgments

This work was supported by the Information and Infrastructure Integrity Initiative of the Pacific Northwest National Laboratory (PNNL). PNNL is operated by Battelle for the U.S. Department of Energy under Contract DE-AC06-76RL01830. The authors are indebted to Chris Brown, a visiting DHS Fellow, for contributions to background research. PNNL Information Release No. PNNL-SA-80437.

10. References

[1] CSO Magazine*, et al.*, "2010 CyberSecurity watch survey - survey results," 2010.

[2] M. Keeney*, et al.*, "Insider threat study: computer system sabotage in critical infrastructure sectors," U.S. Secret Service and Carnegie-Mellon University, Software Engineering Institute, CERT Coordination Center.2005.

[3] F. L. Greitzer and D. A. Frincke, "Combining traditional cyber security audit data with psychosocial data: towards predictive modeling for insider threat," in *Insider Threats in Cyber Security*. vol. 49, C. W. Probst*, et al.*, Eds., Springer US, 2010, pp. 85-114.

[4] L. A. Kramer*, et al.*, "Technological, social, and economic trends that are increasing U.S. vulnerability to insider espionage," Defense Personnel Security Research Center, Monterey, CA TR 05-10, 2005.

[5] E. D. Shaw and L. F. Fischer, "Ten tales of betrayal: the threat to corporate infrastructures by information technology insiders. Report 1 - overview and general observations," Defense Personnel Security Research Center, Monterey, CA TR 05-04, 2005.

[6] E. E. Schultz, "A framework for understanding and predicting insider attacks," *Computers & Security,* vol. 21, pp. 526-531, 2002.

[7] F. L. Greitzer*, et al.*, "Social/ethical issues in predicitve insider threat monitoring," in *Information Assurance and Security Ethics in Complex Systems: Interdisciplinary Perspectives*, M. J. Dark, ed., IGI Global, 2010.

[8] J. Pearl, "Bayesian networks: a model of self-activated memory for evidential reasoning (UCLA Technical Report CSD-850017)," in *7th Conference of the Cognitive Science Society, University of California*, Irvine, CA, 1985, pp. 329-334.

[9] D. Heckerman, "Tutorial on Learning with Bayesian Networks," in *Learning in Graphical Models*, M. I. Jordan, Ed., ed, 1998.

[10] A. J. L. Harris*, et al.*, "Estimating the probability of negative events," *Cognition,* vol. 110, pp. 51-64, 2009.

[11] D. E. Goldberg, *Genetic Algorithms in Search, Optimization and Machine Learning*. Reading, MS: Addison-Wesley, 1989.

[12] J. Holland, "Genetic algorithms," *Scientific American,* vol. 267, pp. 66-72, 1992.

[13] D. E. Rumelhart*, et al.*, "Learning internal representations by error propagation," in *Parallel Distributed Processing: Explorations in the Microstructures of Cognition. 1: Foundations*, D. E. Rumelhart and J. L. McClellands, Eds., ed Cambridge, MA: MIT Press, 1986, pp. 318-362.

[14] P. J. Werbos, "Beyond regression: new tools for prediction and analysis in the behavioral sciences," PhD, Department of Applied Mathematics, Harvard University, Cambridge, MA, 1974.

[15] P. J. Werbos, *The Roots of Backpropagation*. New York: John Wiley & Sons, Inc., 1994.

[16] F. L. Greitzer*, et al.*, "Identifying at-risk employees: a behavioral model for predicting potential insider threats," Pacific Northwest National Laboratory, Richland, WA PNNL-19665, 2010. Available online, http://www.pnl.gov/main/publications/external/technical_reports/PNNL-19665.pdf.

[17] E. Cole and S. Ring, *Insider Threat: Protecting the Enterprise from Sabotage, Spying and Theft*. Rockland, MA: Syngress Publishing, 2006.

[18] D. R. Band*, et al.*, "Comparing insider IT sabotage and espionage: a model-based analysis," Carnegie-Mellon University. Software Engineering Institute. CERT Coordination Center. CMU/SEI-2006-TR-026, 2006.

[19] M. R. Randazzo*, et al.*, "Insider threat study: illicit cyber activity in the banking and financial sector," Carnegie-Mellon University. Software Engineering Institute. CMU/SEI-2004-TR-021, 2004.

[20] S. L. Pfleeger*, et al.*, "Insiders behaving badly: addressing bad actors and their actions," *IEEE Transactions on Information Forensics and Security,* vol. 5, pp. 169-179, 2010.

[21] T. Yarkoni, Personality in 100,000 words: A large-scale analysis of personality and word use among bloggers. *Journal of Research in Personality*, vol. 44, pp. 363-373, 2010.

[22] C. N. DeWall, *et al.*, "Narcissism and implicit attention seeking: Evidence from linguistic analyses, " *Personality and Individual Differences*, vol. 51, pp. 57–62, 2011.

[23] F. L. Greitzer and R. E. Hohimer, "Modeling human behavior to anticipate insider attacks," *Journal of Strategic Security,* vol. 4, pp. 25-48, 2011.