2013 IEEE Security and Privacy Workshops

A Bayesian Network Model for Predicting Insider Threats

Elise T. Axelrad, Paul J. Sticha
Human Resources Research Organization (HumRRO)
Alexandria, VA 22314, USA
{eaxelrad, psticha}@humrro.org

Oliver Brdiczka, Jianqiang Shen
Palo Alto Research Center (PARC)
Palo, Alto, CA 94304, USA
{brdiczka, jshen}@parc.com

***Abstract*—This paper introduces a Bayesian network model for the motivation and psychology of the malicious insider. First, an initial model was developed based on results in the research literature, highlighting critical variables for the prediction of degree of interest in a potentially malicious insider. Second, a survey was conducted to measure these predictive variables in a common sample of normal participants. Third, a structural equation model was constructed based on the original model, updated based on a split-half sample of the empirical survey data and validated against the other half of the dataset. Fourth, the Bayesian network was adjusted in light of the results of the empirical analysis. Fifth, the updated model was used to develop an upper bound on the quality of model predictions of its own simulated data. When empirical data regarding psychological predictors were input to the model, predictions of counterproductive behavior approached the upper bound of model predictiveness.**

Keywords: Insider Threat Detection; Psychological Profiling; Bayesian Network Model

I. INTRODUCTION

Government and corporate organizations face a growing threat of malicious employees who steal confidential information, destroy information systems, or even kill co-workers. These threats often happen without warning and can cause enormous damage. After the fact, however, a pattern or trail can often become evident that could have identified the malicious insider. In most cases, this trail is a combination of suspicious activities (e.g., downloading big files after work hours) paired with a motivational or psychological profile (e.g., by having financial and personal stress) that indicate the desire to commit a malicious act.

This paper defines a Bayesian network model that incorporates psychological variables that indicate degree of interest in a potential malicious insider. We begin by identifying psychological variables hypothesized to characterize a malicious insider, as well as the relationships between these variables. The initial relationships are derived from the psychological literature. We then present a study and data to validate the associations among the measurable variables of our proposed model. The study consists of 486 subjects that responded to a 112 item survey measuring the identified psychological variables. We validate the model by comparing predicted relationships between variables from the surveys to the initial predictions derived from the literature. Our results indicate that the derived relationships are valid, but identify several additional relationships that should be considered. We present a final model based on these results.

II. DEVELOPMENT OF THE BAYESIAN NETWORK

A. Identification and Selection of Predictors

There is a relatively small, but growing body of research that directly addresses insider threats to information systems. For example, work conducted by the Computer Emergency Response Team (CERT) at Carnegie Mellon University [1] has documented the characteristics of known inside attackers in a qualitative manner, but has not identified appropriate base rates for the general population to use for comparisons. In addition, research on counterproductive cyber behaviors [2] surveyed ordinary people about the frequency of these behaviors along with other psychological variables, such as their personalities. Finally, some research has sought to establish the validity of criteria used to determine whether to grant an individual access to sensitive or classified information [3].

From these sources, we developed a list of 83 variables potentially associated with insider threat, and established a ranking by estimating a score indicating the power of each variable to predict degree of interest in a potential malicious insider. The score was based on empirical correlations, where they were available. When data were lacking, we made judgmental estimates by comparing the predictors to similar predictors for which the relationship was known. Two of the authors made these estimates independently and discussed them further to come to consensus for those cases in which we diverged.

We then developed a Bayesian network incorporating variables from the rank ordered list. This model incorporates associations between variables to generate a single score for degree of interest for an individual based on that person's characteristics. The model contains a subset of the original 83 variables, selected based on three criteria: (a) whether the variable can be measured in the workplace; (b) whether the variable has an association value of at least r = .15 with degree of interest; and (c) whether the variable would add new information to the prior set. In other words, if two variables were considered to be highly correlated, it would not be necessary to include both of them.

B. Model Structure

Fig. 1 represents a broad conceptual overview of the Bayesian network model, showing the categories of variables that were selected for inclusion. These are:

- Dynamic environmental stressors including personal life stressors and job stressors
- Static personal characteristics including personality and capability

DOI 10.1109/SPW.2013.35

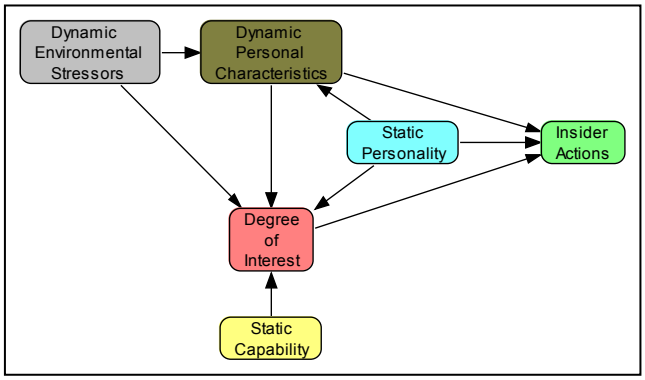

Figure 1. Conceptual model representing the top-level structure of the Bayesian network.

- Dynamic personal characteristics including perceived stress, affect (i.e., hostility), and attitude (i.e., job satisfaction)
- Insider actions (i.e., interpersonal and organizational counterproductive workplace behaviors)
- Degree of interest (i.e., relative risk of insider attack)

C. *Estimation of conditional probabilities*

We drew on many sources for parameters describing the associations between variables in the model, in addition to the results of our initial review. Some research reported pairwise correlations between model variables. Two studies that proved especially useful linked multiple variables in structural equation models (SEM). One of these models [4] links job stress, life stress, and job satisfaction. A second model [5] links personality variables, job satisfaction, and insider actions (i.e., counterproductive workplace behavior). The specific methods used to estimate these parameters are described in the following subsections.

1) *Degree of interest*

The model enables inferences to be made about the relative degree of interest in an individual who has some known and some unknown characteristics. We have eight variables that directly link to this hypothesis variable, and we draw the associations between these variables and degree of interest from the overall score used to rank the predictors. Degree of interest was directly related to excitement seeking, neuroticism, environmental stressors, hostility, and capability; and was inversely related to job satisfaction, agreeableness and conscientiousness. More detail on each of the variables in the equation is presented in the following discussion.

Note that the linking of eight variables to degree of interest creates a large conditional probability table (CPT), given that we use five bins to represent ranges of values for each variable. To simplify the CPTs and improve model efficiency, we partitioned the inputs to degree of interest into two groups of four variables, and created two new variables to represent a weighted subtotal of the inputs for each group. The degree of interest node was then defined as the sum of the two subtotals. Except for some minor differences caused by rounding and the way that variables were discretized, the simplification of the model had no effect on its predictions. In addition, this change reduced the size of the model file by over 90%.

2) *Dynamic Environmental Stressors*

Stressors vary over time and may occur on the job or in other areas of an individual's life. If these stressors lead to hostility or decrease job satisfaction, they may produce a risk of an insider attack. In the model, three environmental stressor variables link to degree of interest in the following ways: (a) The overall environmental stress links directly to degree of interest. (b) Both personal and job stressors, when perceived by an individual may decrease job satisfaction. (c) Finally, stressors may increase the dynamic affect variable, hostility, which, in turn, will increase the degree of interest. We represent all personal stressors in one variable and all job stressors in another. The maximum value of these two is then represented in an environmental stressors variable that is directly associated with degree of interest at r =.15

3) *Static Personality*

The personality of employees affects how they react to stress, their job satisfaction, and their likelihood of engaging in counterproductive behaviors, in addition to being related directly to degree of interest. Also, some people will have the capability to do more damage, especially in the cyber arena, because of their privileged access to information technology, social network connections to hackers and their tools, or their own hacking skill. As such we have introduced personality and capability as static personal variables to consider in assessing relative interest in an individual as a possible inside attacker.

There are five main personality factors: (a) agreeableness, (b) neuroticism, (c) conscientiousness, (d) extraversion, and (e) openness to experience. Each of these factors encompasses several specific facets. For example, agreeableness includes facets of straightforwardness, trust in others, empathy, and others; while extraversion includes warmth, positive affect, and excitement seeking. Our initial review of the literature suggested that the following four personality variables predict degree of interest: (a) excitement seeking, a facet of extraversion; (b) agreeableness; (c) neuroticism; and (d) conscientiousness. Our estimated pairwise correlations between these variables and degree of interest were developed by using a weighted average of values as presented by [2], in which each variable was associated with multiple types of counterproductive cyberbehaviors.

Other related variables include clinical diagnoses found to be present in known inside attackers. These variables include paranoia, depression, narcissism, and sociopathy or antisocial personality disorder. Since clinical diagnoses require clinical interviews, they are unlikely to be collected from all employees. However, clinical diagnoses are substantially correlated with personality [6]. For example, Jakobwitz and Egan [7] report that the "dark triad" of psychopathy, narcissism, and Machiavellianism load on a single factor that is correlated at r = -.69 with agreeableness, one of the personality factors we include in the model. Furthermore, both agreeableness and excitement seeking are correlated with antisocial behavior [6]. Finally, depression and anxiety are facets of the neuroticism personality factor

[8]. Because of the difficulty assessing these clinical variables and their high correlation with personality variables already in the model, we have chosen not to include them as separate variables in the model.

While there is a general agreement that there are five factors of personality, these factors are not statistically independent. Barrett & Roland [9], in a critique of studies of the correlations between personality factors, publish a corrected correlation table of the Big Five factors attributed to a 2007 Hogan Normative Study of 156,614 people. In our original Bayesian network, we did not take into account these intercorrelations between personality factors, except where they were incorporated in the path model predicting counterproductive behavior, to be described under insider actions. Our validation of the Bayesian network addresses the effects of ignoring these relationships on the accuracy of the predictions of the network.

4) Capability

Another factor that affects degree of interest is capability. We are treating this variable as a stable personal variable for the time course assumed in this research, although we know skill does change over time. While capability in and of itself is not a predictor of inside attack (and thus has a low association with insider attack), capability in conjunction with hostile intent increases the extent of the threat. Some crimes such as espionage and fraud may not require much technical skill, such as printing sensitive data onto hard copy or downloading files and handing them off. Others, such as computer sabotage, can be done in a sophisticated way requiring much skill, or can be done using software packages that are available on the internet and require less skill. Those with social connections to hackers do not themselves need skill but have access to that capability. We have modeled capability as a single variable with a r = .15 association with degree of interest.

5) Dynamic Personal Variables: Affect and Attitude

Cases of insider sabotage often include a dynamic buildup of anger and job dissatisfaction after an environmental stressor, such as a demotion [10]. In addition to static personality variables and dynamic stressors, we include affect and attitude variables that have been described as precursors to attack.

The Merriam-Webster dictionary defines affect as "the conscious subjective aspect of an emotion considered apart from bodily changes; also a set of observable manifestations of a subjectively experienced emotion." Angry hostility, the tendency to negatively misperceive others and respond in angry and bitter ways, is the listed measure most associated with degree of interest. It can be measured as hostility using the Positive and Negative Affect Schedule (PANAS, [11][12]). The PANAS can be used as a state measure to assess these values over a finite time span (i.e., today, past few days, past week, past few weeks, past month, past year) or as a measure of an ongoing trait (i.e. "Rate the extent to which you generally experience" the following affect).

Hostility has an association with degree of interest of r = .20. It has a dynamic state aspect, associated with environmental stressors at r = .15 and a static personality aspect that is correlated with neuroticism at r = .40 and agreeableness at r = -.43 [13].

Another category of dynamic personal variable, in addition to affect, is attitude. An attitude is "a positive or negative evaluation of people, objects, and ideas" [14]. Our model captures these variables as perceived life stress, perceived work stress, and job satisfaction. We use r = -.15 as the association between job satisfaction and degree of interest.

In order to link stressors, perceived stress and job satisfaction, we drew on the path model developed by Hendrix et al., [4]. This model links personal stressors, job stressors, perceived life stress, perceived job stress, and job satisfaction. We excluded direct positive influences on job satisfaction and focused on those job stressors that increased job stress, using the path coefficient for the largest of these effects, "work subject to whim of superiors," with a $\beta = .28$. Similarly we focused on the strongest personal life stressor, "home-family relations" with a path coefficient, $\beta = -.52$ as a link to life stress. Life stress was associated with job stress at $\beta = .19$, and finally, job stress was linked to job satisfaction at $\beta = -.11$.

6) Insider Actions: Counterproductive Behavior

Some of the strongest predictors of degree of interest are past insider actions that constitute rule violations, whether of social norms or organizational rules. Robinson & Bennett [15] developed a questionnaire for workplace deviance with a factor structure suggesting two factors: (a) counterproductive behavior towards individuals (CPB-I, e.g. "acted rudely towards someone at work") and (b) counterproductive behavior towards the organization (CPB-O, e.g. "taken property from work without permission").

We modeled counterproductive behavior as two variables, CPB-I and CPB-O. These were modeled as indicators of degree of interest, with associations of r = .28 for CPB-I based on interpersonal conflict and social isolation and r = .35 for CPB-O based on rule breaking.

The counterproductive behavior variables were not only functions of degree of interest, but also of job satisfaction and personality. In order to model these insider actions, we drew on a path model from Mount and colleagues [5] that linked self-report measures of agreeableness, neuroticism, conscientiousness, job satisfaction, CPB-I, and CPB-O.

D. The Bayesian network

Figure 2 depicts the Bayesian network as it was before validation. The primary hypothesis, degree of interest, is at the center of the model, with its score predicting individual and organizational counterproductive behavior, based on input variables of personality, dynamic attitudes and emotions, capability, and stressors.

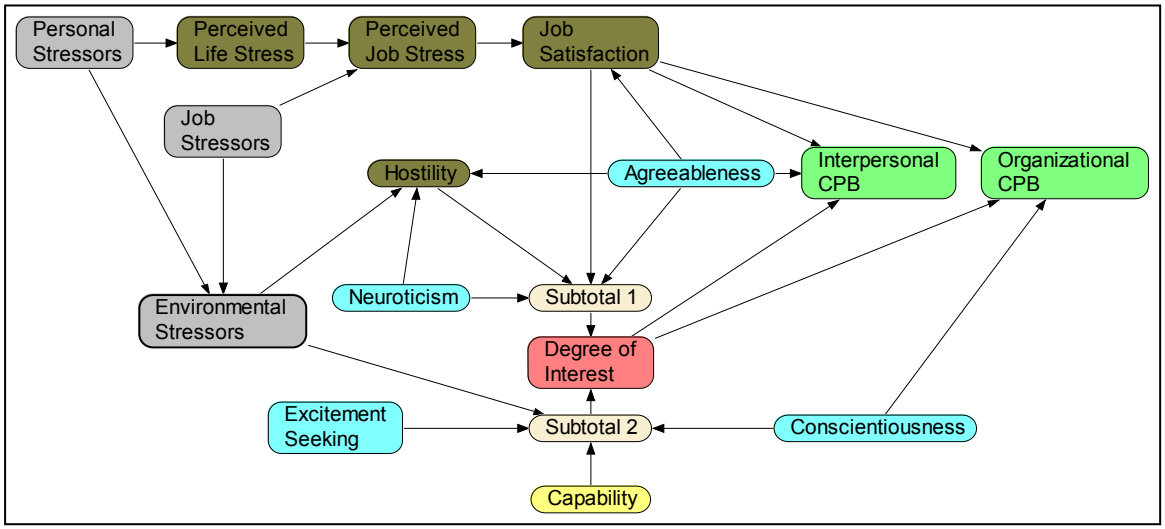

Figure 2. Bayesian network variables and structure.

III. Model Validation

When we developed the original Bayesian network, we estimated conditional probability tables based on associations in the research literature and on expert judgment. The limitation of this approach is that we might find one study that measured the association between, for example, variables A and B, and another that measured the association between variables B and C. These two studies may have measured B differently and certainly collected data using different participants. Finally the association between A and B might be different in the presence of a range of values of C that might differ across studies, whether C was measured or not. The ideal approach is to measure all the relevant variables in one sample using the same method.

In order to validate the Bayesian network, we wanted to measure as many of the variables in the model that could be elicited from participants, to see how these variables were associated in one sample of people when measured with the same questionnaire items. We developed a hypothesized structural equation model (SEM) that incorporated as many of the variables from the Bayesian network as it was feasible to measure. We then split the data in half, randomly assigning cases to a development sample and a holdout sample. We fit the SEM to the first data sample and adjusted the model in light of the results so that we could test it on the holdout sample. We then aggregated the data to fit a final structural model.

A. Survey Data Collection

We conducted a psychological survey using Amazon's Mechanical Turk, a means of employing people to do online tasks and submit results for a small fee. All subjects gave informed consent to participate in our study.

Not all of the variables in the Bayesian network were included in the survey, because the participants were not in a single organization nor would they be willing and/or able to share such information. Specifically, capability (e.g. system administrator privileges or knowledge of malicious software), job and personal stressors (e.g. recent demotion, financial stress), and degree of interest were excluded. Some variables that were in the survey were not in the network because they were not hypothesized to be predictive of degree of interest. These were extraversion, self-assurance, and overall measures of positive and negative affect. We did include a facet of extraversion known to be predictive of criminal activity, excitement seeking. We also included a subset of items from negative affect that measured hostility.

1) Variables Available in the Survey and the Model

We assessed the variables in the survey using measures with established reliability and validity. The variables that appeared in both the survey and the network are shown in the following list, with references to their sources:

- Agreeableness (e.g., straightforwardness, trust in others, and empathy; from the International Personality Item Pool (IPIP [16][17][18]),
- Neuroticism (e.g., anxiety and depression; from IPIP),
- Conscientiousness (e.g., dutifulness and self-discipline; from IPIP),
- Excitement seeking (e.g., recklessness, seeking adventure and danger; facet of extraversion from IPIP),
- Perceived stress (e.g., feeling overwhelmed or a lack of control over important things [19]),
- Hostility (e.g., feeling disgust, anger, and loathing; from Positive and Negative Affect Scale [13]),
- Job Satisfaction (e.g., enthusiastic about work most days [20]),
- Interpersonal Deviance (e.g., counterproductive behaviors directed toward an individual such as publicly embarrassing someone at work [15]), and
- Organizational or Workplace Deviance (e.g., counterproductive behaviors directed against the organization ranging from taking a long break to discussing confidential company information with an unauthorized person [15]).

2) Issues in Measuring Personal and Organizational Deviance and Job Satisfaction

Often when participants are asked to report on their own preferences, attitudes, and behaviors, they do not respond accurately. Sometimes individuals are motivated by impression management or self-deception and sometimes there are lapses in memory. While this is a concern for all variables we measured, we were particularly concerned that the workplace measures of counterproductive behaviors (interpersonal deviance, organizational deviance) and perhaps job satisfaction would not be answered accurately.

We addressed this concern by wording the items as if the participant was answering about other people, making use of the much-replicated false consensus effect [21]. This effect is a bias in which people tend to believe that most other people share their beliefs, attitudes, and preferences. Based on our resulting assumption that people who would do a particular counterproductive behavior x would believe other people would as well, we asked "what proportion of people at work are likely to do x?" rather than "how likely is it that you would do x?" While this approach may avoid the issues with impression management and self-deception, it has the limitation that the individual is answering for others.

B. *Identifying and Removing Unreliable Records*

Participants in online data collection are paid, but they are not supervised in the same way that they would be in a typical experiment. As a result, there is a risk that they would try to game the system somehow to produce data without careful effort. We implemented several methods for identifying and excluding data from participants who answered unreliably (e.g., that similar items weren't answered in similar ways), who responded too rapidly, or who followed simple response patterns (e.g., all 1's). Specifically, we eliminated cases in which any of the following conditions occurred:

- The participant gave inconsistent responses across a set of matched items;
- The total time to complete the survey was less than 2 minutes (the median response time was 7:15);
- The participant responded with a long string of a single response alternative;
- The participant gave responses with low variance;
- The participant did not complete the entire survey.

Overall, data from 89 participants were removed producing a total dataset with 486 observations. Once the data were collected and cleaned of inaccurate or unreliable records, we split the dataset into two parts, each with 243 observations; one of these parts was used for model development, while the other was used for model validation.

C. *Developing Structural Equation Models*

We fitted a structural equation model to the development sample using maximum likelihood estimation. We set up the personality variables and perceived stress as predictor variables, with a fixed variance of 1.00. We fitted path coefficients where there were links in our Bayesian network, excluding the unmeasured variables. When we ran the model using LISREL software, the results suggested model modifications that would improve the fit. Selecting from among these suggestions, we added two more paths: One was a correlation between the counterproductive behavior measures (personal and organizational deviance) that was captured in the Bayesian network through the degree of interest variable and was described in the original research literature about the measure ([22], r = .68, n=226). The other was a path between excitement seeking and hostility. In addition, the model fit correlations among the personality variables and perceived stress. When we conducted this fit in the test sample, we found a new set of model parameters, similar to those in the development sample. Table I presents the model path coefficients as estimated in the development sample, the test sample, and the literature-based estimates we had used in the Bayesian network. The parameter estimates replicated in the two datasets suggesting a plausible model.

Since the path coefficients were similar in the development and test sample, we fit the model to the full data sample. The resulting model is shown in Fig. 3.

TABLE I. COMPARISON OF MODEL PATH COEFFICIENT ESTIMATES IN THE DEVELOPMENT SAMPLE, TEST SAMPLE, AND LITERATURE

Link		Develop-ment	Holdout	Full Sample	Literature Based
From	To				
Neurotc	Hostile	0.52	0.52	0.51	0.40
ExctSeek	Hostile	0.29	0.12	0.20	N/A
Agreebl	Hostile	-0.36	-0.43	-0.41	-0.43
Agreebl	JobSat	0.15	0.06	0.10	0.36
PerStress	JobSat	-0.35	-0.33	-0.35	-0.11
Conscien	OrgDev	-0.19	-0.11	-0.15	-0.52
Agreebl	PersDev	-0.19	-0.07	-0.13	-0.34
JobSat	PersDev	-0.27	-0.21	-0.24	-0.40
JobSat	OrgDev	-0.29	-0.24	-0.26	-0.41

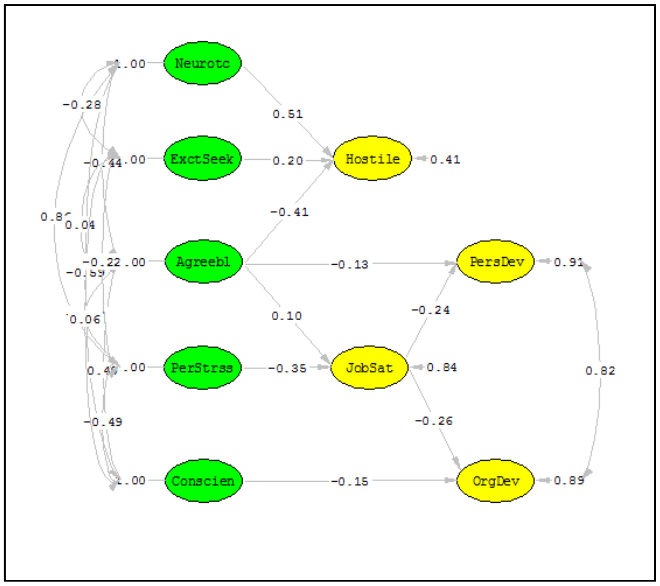

Figure 3. Structural equation model for the full sample of data.

The model fit statistics showed a reasonable but not perfect fit, with a comparative fit index (CFI) of 0.95 and an RMSEA of 0.064. With a sample size of n=486, the chi-square was highly sensitive at Chi-Square = 7241.35, df = 3060, p=0.0 and a p-value test of close fit (RMSEA <0.05) of 0.00. The standardized RMR was 0.086 and the adjusted Goodness of Fit index (AGFI) was 0.66.

As a reviewer pointed out, a k-fold validation might have provided additional confidence about the model. However this approach was not chosen because of the additional time required to estimate multiple SEMs.

D. Incorporating Results into the Bayesian Network

The results of the SEM produced a structure that was similar in most respects to the Bayesian network. However, there were several differences in the model structure and parameter values between the two models. We changed the structure of the Bayesian network to reflect the following two relationships that were identified in the SEM: (a) a link from excitement seeking to hostility with conditional probabilities reflecting the path coefficient for the SEM, and (b) correlations among personality variables and a correlation between personality variables and perceived stress.

In order to represent the correlations between personality variables and perceived stress as simply as possible, we began by conducting a principal component analysis of the correlation matrix of the personality and perceived stress variables. This analysis yielded two factors, which accounted for 71% of the total variance. We rotated these factors using a varimax rotation. The first of these factors had high loadings from neuroticism, perceived stress, agreeableness, and conscientiousness, with the loading for the first two variables having the opposite sign as the loading for the last two. The second factor had a high loading for excitement seeking.

We created a normally distributed node to represent the first factor. This node was linked to the four nodes included in the factor. The CPT for each of the linked personality and stress variables reflected its loading in the rotated factor matrix. Since the second factor contained only a single variable, we did not have to make any changes to the Bayesian network to represent it.

Finally, we updated the CPTs for other nodes that corresponded to the latent variables in the SEM, using the appropriate path coefficients to represent the relationship between the value of a node and the values of its parents. While we believe that it would be preferable to combine the information from the SEM analysis of the survey and the original information from the research literature, aggregating the information would be difficult, given the structural changes that were made in the model. Consequently, where there were results from the SEM, the path coefficients replaced the partial correlations that were used to estimate the original model. The resulting model is shown in Figure 4.

E. Testing the Bayesian Network

The primary hypothesis variable for the Bayesian network was the degree of interest. Since this variable was not included in the survey, we could not test the accuracy of prediction directly related to the hypothesis variable. However, both the survey and the Bayesian network include two indicators of degree of interest, interpersonal counterproductive behavior and organizational counterproductive behavior. We used these variables as criteria to examine the extent to which the changes in the Bayesian network affected its ability to predict counterproductive behavior.

The following strategy was used to test the predictions of the Bayesian network. The network processed a number of cases in which the values of variables included in the survey, except for the two measures of counterproductive behavior, were specified. These values were entered into the network as findings (or evidence), and the network then predicted the probability of counterproductive behavior based on this evidence. The predicted values were then compared to the actual values to assess the correctness of the predictions. For the original and then the revised Bayesian network, we conducted this analysis for a set of cases simulated using the Bayesian network to get an upper bound on the possible accuracy of model prediction and repeated the analysis using actual cases from the survey data set.

Cases came from two sources, (a) simulated cases that were generated by the Bayesian network itself, and (b) empirical cases based on responses to the survey. The empirical cases were drawn directly from the survey measures, but were normalized to have means and standard deviations that corresponded to the comparable variables in the Bayesian network. The simulated cases provided a baseline against which the quality of the predictions of empirical cases was assessed.

Table II shows three measures of prediction accuracy: error rate, logarithmic loss, and quadratic loss. The two types of counterproductive behavior had five levels, ranging from the lowest indicating a reported low incidence of counterproductive behaviors to the highest indicating a high incidence. The error rate indicates the percentage of cases for which the model-predicted value of the counterproductive behavior variable was different from the actual (either empirical or simulated) value of the variable. Although the network predicts a distribution of probabilities across the levels of the variable, the model-predicted level of the variable is the level with the highest estimated probability. While the error rates, which range from 48% to 67%, appear high for all conditions, this result reflects the number of response categories and the moderate magnitude of the relationships between variables. Since the error rate associated with random prediction is 80% (20% probability for each of 5 levels), the error rate values obtained for the network are a substantial improvement over random prediction.

The error rates based on data simulated from the network indicate the level of prediction that is possible from the network when it accurately represents the conditional relationships between variables. Comparing the error rates for the simulated data for the original and revised networks suggests that the revised network makes somewhat weaker (or less extreme) predictions than the original network,

TABLE II. COMPARISON OF MODEL PREDICTION ACCURACY BETWEEN ORIGINAL AND REVISED BAYESIAN NETWORKS

Variable	Original Network		Revised Network	
	Simulated Data	Empirical Data	Simulated Data	Empirical Data
Interpersonal Counterproductive Behavior				
Error Rate	54.94%	66.46%	60.08%	64.81%
Logarithmic Loss	1.207	1.427	1.302	1.344
Quadratic Loss	0.657	0.738	0.689	0.717
Organizational Counterproductive Behavior				
Error Rate	47.74%	66.87%	55.35%	61.93%
Logarithmic Loss	1.069	1.611	1.294	1.334
Quadratic Loss	0.599	0.792	0.682	0.709

leading to its higher error rate. However, the revised network makes fewer errors than the original network when it receives the empirical data from the survey.

The two loss functions provide more sensitive measures of the accuracy of the predictions of the Bayesian network, in that they consider the entire predicted probability distribution over the levels of the criterion variables, rather than just the level with the greatest probability. Both of these functions penalize a prediction when the level of counterproductive behavior that actually occurs in a simulated or empirical case is predicted with a low probability. The possible values of a logarithmic loss function range between zero (i.e., the prediction is perfect) and infinity, while the quadratic loss function values range from zero to two. Our model shows logarithmic loss values between 1 and 1.7, (i.e., relatively close to 0) and quadratic loss values ranging between 0.6 and 0.8 (closer to 0 than to 2).

Both of these loss functions provide a similar picture of the relative accuracy of the original network and the revised network in predicting counterproductive behavior, and a picture that is consistent with that provided by the error rate. Specifically, the revised network places somewhat lower probabilities on the correct predictions of the level of counterproductive behavior than the original network, when evidence is simulated from the network. However, the revised network makes better predictions from the empirical survey data than the original network. The results for the two forms of counterproductive behavior, interpersonal and organizational, are similar.

We developed confusion matrices comparing the predicted and actual levels of counterproductive behavior from the empirical survey responses, in order to provide additional detail about the predictive validity of the original and revised networks. We found only a small number of actual cases in the lowest level (none for interpersonal, and 14 out of 486 for organizational counterproductive behavior). This reflects the fact that the survey measures of interpersonal and organizational deviance, which correspond to the two counterproductive behavior measures contained in the network, are positively skewed. People do not report many such behaviors so most people respond at the low end of the scale. Consequently, 2.5 standard deviations below average (the upper boundary of the lowest category) is well below the minimum value in the sample. Since the interpersonal and organizational deviance scores from the surveys are normalized based on their mean and standard deviation, a positively skewed distribution will produce normalized variables that do not deviate sufficiently from the mean in the negative direction. Second, the original network makes more extreme predictions (both at the low and high end) than the revised network. This suggests that the correlations in the revised model are somewhat weaker than those in the original model. It may also be a result of the intercorrelations between personality variables that were included in the revised network.

IV. DISCUSSION

We started with a Bayesian network developed from associations among psychological variables described in the research literature. Data were collected to measure a subset

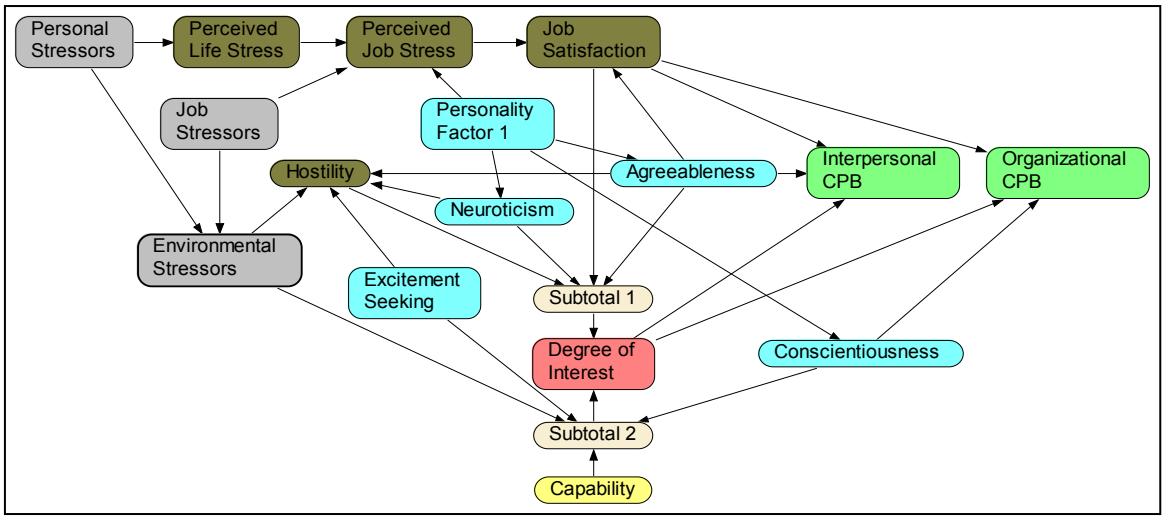

Figure 4. Revised Bayesian network based on results of SEM.

of these variables in a sample of participants. We split the sample and tested a structural equation model that was designed to represent a portion of the Bayesian network. We found that the estimated path coefficients in the structural model were similar but not exactly the same as those derived from the research literature.

The results suggested some changes to the original model that would improve model fit. Based on the results of that fit, we implemented those adjustments and tested the adjusted model on a holdout sample. The fit of the revised model was reasonable so we adjusted the original Bayesian network in the same manner.

We tested the Bayesian network-based predictions of counterproductive behavior by fixing predictive variables to their values in a simulated dataset (and predicting simulated counterproductive behavior) and then tested it again in the empirical dataset. We found that the revised model predicted the simulated data less well than the original model, most likely because the revised fit lowered some of the associations between variables. Nevertheless, the revised model predicted the empirical data better than the original model. Three limitations on the predictiveness of the model include the low associations between variables in the environment (i.e., not all people who fit a predictive profile commit crimes), the proxy measure we used for counterproductive behaviors, and the fact that counterproductive behaviors are rare events. Nevertheless, with most of the structure and parameters from the original Bayesian network model maintained in the updated model, we believe we have provided a reasonable validation.

ACKNOWLEDGMENT

The authors gratefully acknowledge support for this work from DARPA through the ADAMS (Anomaly Detection At Multiple Scales) program funded project GLAD-PC (Graph Learning for Anomaly Detection using Psychological Context). Any opinions, findings, and conclusions or recommendations in this material are those of the authors and do not necessarily reflect the views of the government funding agencies.

REFERENCES

[1] S.R. Band, D.M. Cappelli, L.F. Fischer, A.P. Moore, E.D. Shaw, and R.F. Trzeciak, Comparing Insider IT Sabotage and Espionage: A Model-Based Analysis (Technical Report CMU/SEI-2006-TR-026; ESC-TR-2006-091). CERT Program, Pittsburgh,PA: Carnegie Mellon University Software Engineering Institute, 2006.

[2] S.S. Russell, M.J. Cullen, M.J. Bosshardt, S.E. Juraska, A.L. Stellmack, E.E. Duehr, and K.R. Jeansonne, Cyber Behavior and Personnel Security (Institute Report #661), Minneapolis, MN: Personnel Decisions Research Institutes, Inc, 2009.

[3] Adjudicative Guidelines for Determining Eligibility for Access to Classified Information. Adjudicative Desk Reference: Executive Orders and Regulations, 2005. Downloaded from http://www.dhra.mil/perserec/adr/adjguidelines/adjguidframeset.htm on 2/11/2013.

[4] W.H. Hendrix, N.K. Ovalle, and R.G. Troxler, "Behavioral and physiological consequences of stress and its antecedent factors," Journal of Applied Psychology, vol. 70(1), 1985, pp. 188-201.

[5] M. Mount, R. Ilies, and E. Johnson, "Relationship of personality traits and counterproductive work behaviors: The mediating effects of job satisfaction," Personnel Psychology, vol. 59, 2006, pp. 591-622.

[6] B.P. O'Connor and J.A. Dyce, "Tests of general and specific models of personality disorder configuration," in Personality Disorders and the Five-Factor Model of Personality, P. T. Costa and T. A. Widiger, Eds. Washington, DC: American Psychological Association, 2002, pp. 223-246.

[7] S. Jakobwitz and V. Egan, "The 'dark triad' of psychopathy and normal personality traits," Personality and Individual Differences, vol. 40, 2006, pp. 331 – 339.

[8] P.T. Costa and R.R. McCrae, Revised NEO Personality Inventory and NEO Five-Factor Inventory professional manual. Odessa, FL: Psychological Assessment Resources, 1992.

[9] P. Barrett and P. Rolland, "The meta-analytic correlation between two Big Five factors: Something is not quite right in the woodshed," 2009. Retrieved on 1/12/2012 from http://www.pbarrett.net/stratpapers/metacorr.pdf.

[10] D.M. Cappelli, A. Moore, and R. Trzeciak, The CERT Guide to Insider Threats: How to Prevent, Detect, and Respond to Information Technology Crimes (Theft, Sabotage, Fraud), SEI Series in Software Engineering. Upper Saddle River, NJ: Pearson Education, Inc, 2012.

[11] D. Watson, L.A. Clark, and A. Tellegen, "Development and validation of brief measures of positive and negative affect: The PANAS scales," Journal of Personality and Social Psychology, vol. 54, 1988, pp. 1063-1070.

[12] D. Watson and L.A. Clark, "The PANAS-X: Manual for the Positive and Negative Affect Schedule – Expanded Form," 1994. Retrieved 1/11/12 from http://www.psychology.uiowa.edu/faculty/clark/panas-x.pdf

[13] D. Watson and L.A. Clark, "On traits and temperament: General and specific factors of emotional experience and their relation to the five-factor model," Journal of Personality, vol. 60(2), 1992, pp. 441-476.

[14] R. Gerrig and P.G. Zimbardo, Psychology and life (19th ed.), 2010. Boston, MA: Allyn & Bacon

[15] S. Robinson & R. Bennett, "A typology of deviant workplace behaviors: a multidimensional scaling study," Academy of Management Journal, vol. 38, 1995, pp. 555-572.

[16] T. Buchanan, J.A. Johnson, and L.R. Goldberg, "Implementing a five-factor personality inventory for use on the Internet," European Journal of Psychological Assessment, vol. 21, 2005, pp. 115-127.

[17] L.R. Goldberg, "International personality item pool: A scientific collaboratory for the development of advanced measures of personality traits and other individual differences," 2005. Retrieved January 29, 2013, from the International Personality Item Pool Web site: http://ipip.ori.org

[18] L.R. Goldberg, J.A. Johnson, H.W. Eber, R. Hogan, M.C. Ashton, C.R. Cloninger, and H.C. Gough, "The International Personality Item Pool and the future of public-domain personality measures," Journal of Research in Personality, vol. 40, 2006, pp. 84-96.

[19] S. Cohen, T. Kamarck, and R. Mermelstein, "A global measure of perceived stress," Journal of Health and Social Behavior, vol. 24(4), 1983, pp. 385-396.

[20] A.H. Brayfield and H.F. Rothe, "An index of job satisfaction," Journal of Applied Psychology, vol. 35, 1951, pp. 307-311. Five item subset retrieved from http://www.rotman.utoronto.ca/~scote/questionnaires.pdf.

[21] L. Ross, "The false consensus effect: An egocentric bias in social perception and attribution processes," Journal of Experimental Social Psychology, vol. 13(3), 1977, pp. 279–301.

[22] R.J. Bennett and S.L. Robinson. "Development of a measure of workplace deviance," Journal of Applied Psychology, vol. 85(3), 2000, pp. 349-360.