Methods and Metrics for Evaluating Analytic Insider Threat Tools

Frank L. Greitzer
PsyberAnalytix
Richland, WA USA
Frank@PsyberAnalytix.com

Thomas A. Ferryman
Astute Analytics
Richland, WA USA
tomferryman1@gmail.com

***Abstract*— The insider threat is a prime security concern for government and industry organizations. As insider threat programs come into operational practice, there is a continuing need to assess the effectiveness of tools, methods, and data sources, which enables continual process improvement. This is particularly challenging in operational environments, where the actual number of malicious insiders in a study sample is not known. The present paper addresses the design of evaluation strategies and associated measures of effectiveness; several quantitative/statistical significance test approaches are described with examples, and a new measure, the *Enrichment Ratio*, is proposed and described as a means of assessing the impact of proposed tools on the organization's operations.**

Keywords—insider threat; evaluation; validation; metrics; assessment

I. INTRODUCTION

The insider threat refers to harmful acts that trusted individuals might carry out, causing harm to the organization or its personnel, 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 insider threat problem covers a broad range of activities, with policy violation at one end of the continuum of abuses and espionage/sabotage at the other. A recent book by Carnegie-Mellon's CERT program [1] provides a comprehensive reference, discussion of cases, and description of best practices in the prevention, detection, and response to IT insider crimes. An informative review of research approaches and challenges may be found in the IATAC SOAR report [2]. A framework for discussing best practices is provided in reference [3].

Currently, no single threat assessment technique gives a complete picture of the insider threat problem. Approaches to insider threat detection vary based on the types of data monitored as well as the nature of the analytic method employed. Typical monitoring approaches in current use incorporate host/network-based monitoring to derive forensic measures including external threat/defense-oriented appliances such as Intrusion Detection or Prevention Systems and Data Leak Detection/Prevention Systems. Several researchers have argued that a comprehensive analytic approach is needed that incorporates monitoring and analysis of a variety of data from cyber monitoring of computer/network activity to behavioral observations and human resources data [4] [5] [6].

There are several technical approaches to analysis of monitored data aimed at detecting or predicting threats. A recent review [7] describes broad technical approaches to intrusion detection (including insider threats) that may be characterized in terms of threshold, anomaly, rule-based, and model-based methods [8]. Threshold detection is essentially summary statistics (such as counting events and setting off an alarm when a threshold is exceeded). Anomaly detection is based on identifying events or behaviors that are statistical outliers; a challenge is to effectively combat the strategy of insiders to work below the statistical threshold of tolerance and, over time, train systems to recognize increasingly abnormal behavior patterns as normal. Rule- or signature-based methods are limited to work within the bounds of the defined signature database; variations of known signatures are easily created to thwart such misuse-detectors, and completely novel attacks will nearly always be missed. Model-based methods seek to recognize attack scenarios at a higher level of abstraction than the other approaches, which largely focus on audit records exclusively as data sources. Regardless of the specific analytic and data monitoring approaches employed, there is a critical need to define and adopt rigorous means of evaluating the effectiveness of proposed solutions. This need is shared by the research community as well as operational users who must decide or choose among proposed technical solutions in the marketplace. The purpose of the present paper is to provide an overview and discussion of methods and metrics for evaluating analytic insider threat tools and approaches.

II. METHODS

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.cyber.st.dhs.gov/docs/IRC_Hard_Problem_List.pdf). Another challenge is the lack of appropriate data and "ground truth" for evaluating detection performance. The challenge is exacerbated because there is a large degree of overlap between observable or measurable behaviors associated with normal versus malicious activities, and a related statistical difficulty in 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 (when

DOI 10.1109/SPW.2013.34

ground truth is known), but due to the nature of the problem, applicable cases are rare. Lacking ground truth data, evaluation methods often adopt strategies to test tools or models against expert judgments. 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.

The challenges are great, but the need is such that the research community must increase its focus on evaluation methods and metrics. To facilitate the discussion, this section describes three general evaluation strategies: testing against expert judgments, injection testing, and testing performance against known outcomes.

A. Testing Against Expert Judgments

While an empirical test is the ultimate aim, other evaluation approaches can be used to test predictions of a model—specifically, to measure the 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. Compare the experts' judgments with the other experts. The inter-expert agreement is a factor in assessing the model's effectiveness.

An example of this approach was described in a research project investigating behavioral models of insider threat [6]. The objective of the study was to validate a psychosocial component of an insider threat model that uses behavioral data or observations of a number of behavioral indicators such as disgruntlement, etc. (details are described in [5] and [6]). The evaluation study solicited judgments from expert evaluators who examined the same observables used by the model(s). The expert judgments were obtained by asking a collection of ten experts to rate insider threat risk of 24 scenarios that differed in the quantity and severity of behavioral indicators (risk judgments were provided on a 0-10 scale and then normalized to a 0-1 scale). Kendall's w nonparametric coefficient of concordance measuring inter-rater agreement was 0.707, where 0 indicates no agreement and 1, perfect agreement. The highly significant coefficient of concordance suggests there is a high level of agreement among the raters and the agreement is statistically significant ($p < 0.001$). The human expert judgments were then compared with the outputs of several alternative threat models. For example, one model had been developed based on fitting Bayesian weights and probabilities for the psychosocial indicators to the judgments provided by two HR experts. A more simplistic model was developed simply by counting the number of behavioral indicators that were observed, regardless of the possible weights or severity of indicators. The predictions of the models could be plotted against the expert judgment data in a scatterplot; results clearly showed that the counting model was inadequate in describing the process used by experts in assessing psychosocial threat, while the Bayesian model performed adequately: The counting model yielded a R^2 of 0.26 compared to $R^2 = 0.60$ for the Bayesian model (the scatterplot for the Bayesian model is shown in Figure 1).

Even though this method lacked the appeal of testing models against ground truth data, the method was useful and informative in providing statistically significant descriptions of the relative weightings used by human experts in assessing behavioral indicators of insider threat, yielding insight into this process and modeling/statistical analysis methods that could potentially be used to develop analytic risk measures based on the behavioral indicators that could contribute to an overall comprehensive analytic tool.

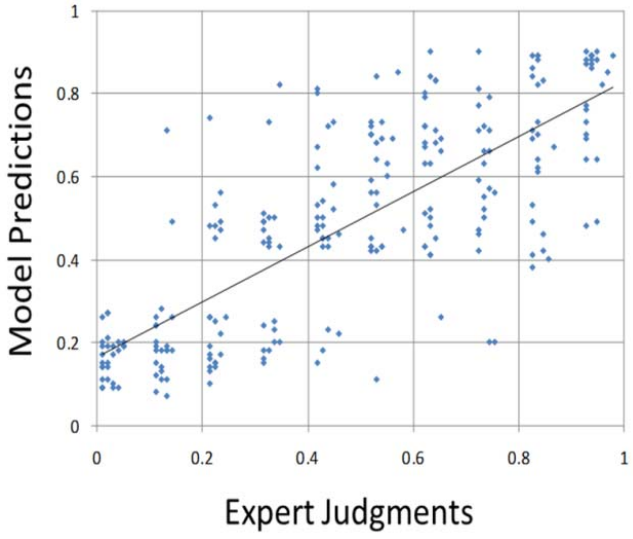

Figure 1. Bayesian prediction of 24 unique cases for a total of 240 test cases

B. Injection Testing

As noted earlier, a critical challenge for insider threat research is lack of actual data that includes ground truth data. In some cases, one might acquire real data, but for privacy reasons, there is no attribution of any individuals relating to

abuses or offenses—i.e., there is no ground truth. The data may contain insider threats, but these are not identified or knowable to the researcher. There have been several examples of such cases in our research on psychosocial indicators of "at-risk" employees, for which we used corpuses of email data to analyze word-use and detect certain targeted personality traits of interest [9] [10].

As a demonstration and proof-of-concept, we applied a word analysis to an existing email corpus (a proprietary dataset representing 167 senders) to determine if it can discern individual word use patterns associated with personality traits among the senders. The outliers may, according to our theory, represent an elevated risk for insider threat. Unfortunately, we did not know the "ground truth" regarding the personalities or behaviors of the individuals who sent these emails (however, we know that some of the individuals represented in the dataset were terminated). To truly validate our methodology, we would need both email samples and objective personality assessments of the senders (or ground truth from relevant employee records).

After pre-processing that removed text not attributable to senders, the dataset comprised approximately 5.25 million words. The mean number of words per individual was about 31,000. Analysis of word use, using the Linguistic Inquiry and Word Count (LIWC) program [11], yielded a tab delimited output file with a row of raw LIWC category frequencies for each sender. The selected word categories were then grouped by their corresponding personality factor. The number of words in each of the word groups was computed for each sender and the counts were converted to a percentage of the total word count for each sender. These percentages were standardized using $z_{ij} = (x_{ij} - m_j) / s_j$, where x_{ij} denotes the percentage for word group ij for sender i; mj denotes the mean of the percentages for word group j over all senders and sj denotes the standard deviation of the percentages for word group j over all senders. The resulting standardized distributions of Neuroticism, Agreeableness, and Conscientiousness scores were derived. Lacking "ground truth" for the senders of these emails, we introduced into the dataset text samples from individuals for whom we had some ground truth information. In particular, to consider the question of whether our methodology could detect an individual who was more like the typical "insider," we added text samples that we were able to obtain for Aldrich Ames, a late 20th century double agent who was convicted of espionage; Benedict Arnold, America's first traitor; and Anna Chapman, the Russian "illegal" who was caught in 2010. We also included samples from other known criminals, though not specifically noted for espionage: Anders Breivik, Adolph Hitler, and Ted Kaczynski. Some of these samples had considerably lower word counts than individuals in the email corpus (viz., for Ames we had a sample of 817 words; 4574 words for Arnold; and only 82 words for Chapman). Adding these "target" individuals to the corpus represents an attempt to assess the ability of the word analysis method to identify likely POIs. The Z-scores for the distribution of (Lack of) Agreeableness scores is shown in Figure 2 for the 167 senders and the six "injected" target individuals (shown as solid-filled histogram bars).

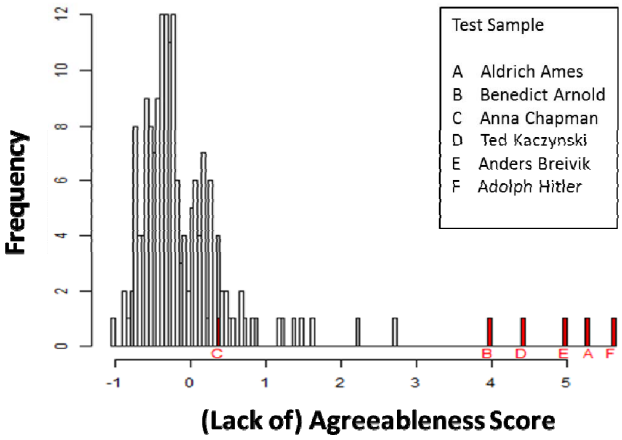

Figure 2. Distribution of Agreeableness Scores (after [9]).

It is evident that individuals in the target group are readily discriminated from the main corpus; they are ranked in the top six highest scores on Neuroticism, in the top seven highest scores in (Lack of) Conscientiousness, and five of the six target individuals are at the top of the (Lack of) Agreeableness scale. Mann-Whitney U-tests conducted for each of the personality trait distributions supported what was obvious in the graphs, with each test identifying highly significant differences in the ranks of the target group compared to the main corpus ($p<.0001$ in all cases).

To statistically identify *individual* outliers in the distributions, we adopted a conservative, distribution-free statistical test that does not depend on assumptions of normally distributed data, based on Chebyshev's inequality, which states that the probability of a random variable (τ) exceeding any real value $T > 0$ is

$$P\left(\left|\frac{\tau - \mu_\tau}{\sigma_\tau}\right| \geq T\right) \leq \frac{1}{T^2} \quad (1)$$

To use the Chebyshev theorem, we must make the assumption that the population variance is finite and not zero. We are confident in this assumption. Additionally, we are confident that the distribution is unimodal, which allows the use of a refinement proven by Vysochanskij and Petunin [12] that allows Chebyshev's original inequality to be tightened by multiplying the right hand side by (4/9). Using the standard score Z_τ in this expression, this yields the following variant of Chebyshev's inequality:

$$P\left(|Z_\tau| \geq T\right) \leq \frac{4}{9T^2} \quad (2)$$

For a given significance level P, a critical value of Z_τ is obtained by solving the equation for Z_τ, yielding:

$$|Z_{crit}| \geq \sqrt{\frac{4}{9P}} \quad (3)$$

At the 5% significance level, the critical value for Z is $\pm$ 2.98. Applying this criterion to the computed Z statistics for our known criminals, the test determined that Aldrich Ames, Benedict Arnold, Ted Kaczynski and Adolf Hitler are outliers in the population with significant scores in all three personality trait categories. Anders Breivik is a significant outlier in two of the personality trait categories.

Reference [9] describes additional analysis methods aimed at identifying *outliers* using significance testing of Mahalanobis distance [13] that should also be consulted for useful analyses. The use of Mahalanobis distance enables derivation of a multivariate metric that combines multiple dimensions (for example, the personality dimensions of Neuroticism, Agreeableness, and Conscientiousness that were studied in the example in Section C; more generally, distributions of various types might be combined into a multivariate distribution that reflects behavioral indicators, personality indicators, and cyber monitoring risk scores). The distance measure for Mahalanobis distance computes the distance from the center of mass of the multi-dimensional multivariate distribution. This metric differs from Euclidean distance in that it takes into account the correlations of the data set and it is scale-invariant. It is particularly useful in cases where the multivariate data are spread out from their center of mass in a non-spherical distribution (e.g., ellipsoidal, as would be the case if the components of the data had different variances—this was indeed the case for the example described in Section B; the multivariate distribution of the three personality trait measures was elongated). Statistical significance tests (such as Mann-Whitney U test) may also be used to test for differences among the specified test groups on the Mahalanobis measure.

The analysis demonstrated here may be examined in more detail in the original source [9]; the point is that the injection methodology and accompanying analysis provides a statistical method for applying statistical significance tests to outcomes of insider threat detection tools. The example was applied to an injection testing methodology that used a test set for injection that was rather different from the population corpus: There were obvious differences in format, culture/temporal era in which samples were generated, and the nature of the perpetrators, to name a few). But the evaluation method is of primary concern for this exposition; it is important to point out that the method is appropriate for cases that contain ground truth for the general population and identified "persons of interest" (POI) in the population. For this preferred case of known outcomes, additional analytic/assessment methods are discussed next.

C. Testing Performance Against Known Outcomes

The most rigorous form of evaluation of a predictive model is to test the predictions against a set of real cases that include POIs/perpetrators (when ground truth is known). The analytic methods for validation described in Section II *B* above apply as well to this case—particularly the nonparametric Chebychev analysis of distributions. When ground truth is available (for known perpetrators or POIs), the assessment of a model or tool should take into consideration not only detection rates but also "false positive" rates (the probability of incorrectly identifying someone as a POI). A general problem with cybersecurity threat detection tools is that detection performance comes with a high false positive rate, which places a high processing load on human analysts and cybersecurity personnel to investigate the large number of leads that are generated. Fortunately, a large body of human performance/signal detection and classification research may be tapped to apply analytic methods of Signal Detection Theory [14] to this problem.

A mathematical framework for describing and studying decisions that are made in uncertain situations, Signal Detection Theory is well-suited for assessing and comparing performance of a detection system under differing conditions. Detection, classification, memory, and even decision making (e.g., diagnostic) performance can be described in terms of four performance scores or probabilities, as shown in a table that is sometimes referred to as a confusion matrix (see Figure 3). The rows correspond to the system's response (e.g., signal present versus signal not present). The columns reflect the true state (SIGNAL versus NO-SIGNAL). In the present context, SIGNAL corresponds to "Person of Interest is Present" and NO-SIGNAL corresponds to "Person of Interest is Not Present." Probabilities or proportions of the system's responses in each cell of the table are referred to using labels such as "true positive" (response = signal present when true state = SIGNAL); "false positive" (response = signal present when true state = NO-SIGNAL); etc. From data collected based on the output of the system (e.g., "yes"/"no" responses and/or confidence ratings about the presence/absence of the signal), analyses derived from Signal Detection Theory may be employed to discern levels of performance. Most relevant are the probabilities of true positive responses and false positive responses, often called "hits" and "false alarms," respectively.

	TRUE STATE	
System Response	**SIGNAL**	**NO SIGNAL**
"Signal Present"	True Positive ("hit")	False Positive ("false alarm")
"Signal Not Present"	False Negative ("miss")	True Negative

Figure 3. Terms in a Confusion Matrix.

A Receiver Operating Characteristic (ROC) curve plots the hit rates against the false alarm rates (see Figure 4). For so-called "yes/no" experiments in which the response is that the signal is either present or not, hit and false alarm probabilities generate a single point on the ROC curve. This point reflects the decision or cut off point of the decision maker. For tests that allow the system to indicate its confidence that a signal is present, multiple points on the ROC curve may be generated [15] (see also textbooks on mathematical psychology, e.g., [16]). The best possible prediction method (perfect classification) would yield a point in the upper left corner or coordinate (0,1) of the ROC space. Because one minus the hit rate is the "false negative" rate, the ROC curve also can be

viewed as the tradeoff between the false negative and false positive rates, for every possible decision cut-off. Good performance is characterized by low false positive rates (false alarms) and low false negative rates (i.e., high hit rates) across a reasonable range of cut off values. Therefore, desirable performance is reflected in ROC curves that are furthest from the (lower left to upper right) diagonal, approaching the (0,1) coordinate. The diagonal line is the expected ROC curve that would be obtained for random performance. For example, if an observer said yes randomly 80 percent of the time no matter what, the hit rate when the signal is actually present would be 80% and the false alarm rate would also be 80% yielding the point (0.80, 0.80) on the ROC curve.

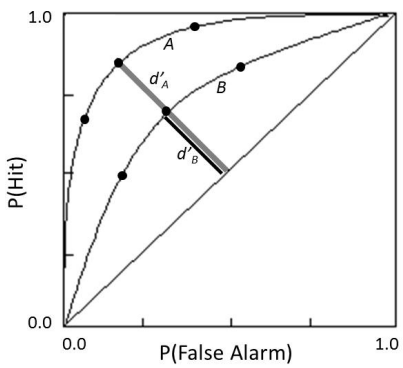

Figure 4. The ROC curve represents decision making performance under uncertainty

In addition, the theory is used to separate two important characteristics of the "receiver": its *sensitivity* and its *bias*. The sensitivity measure (d') is defined as the distance between the "signal" and "signal + noise" distributions in standard error units, i.e., $d' = Z_{signal+noise} - Z_{signal}$; it is also equivalent to the distance from the diagonal as well as the area under the ROC curve [14]. The bias measure is reflected in the position of a point along the ROC curve, and performance may be influenced by manipulating the bias of the observer by such means as varying the reward and/or punishment for correct responses and errors. Sensitivity, on the other hand, is not manipulated by payoff, and therefore it represents the system's level of capability: If system *A* exhibits a higher d' compared to system *B*, then when they are tested under similar conditions, system *A* should be expected to perform better. In this sense, d' provides a means of assessing effectiveness of tools, but there are some practical difficulties: (i) Ground truth must be available. One cannot compute "hit" and "false alarm" rates without having definitive outcome data. (ii) For prediction problems, there is a difficulty in scoring outcomes and therefore assessing hit and false alarm probabilities.

III. PREDICTION CHALLENGES

There is much justification for pursuing the development and validation of a predictive system, as opposed to a detection system that will inevitably relegate defenders to a forensic strategy [6]. Among the most compelling reasons is the finding reported by Shaw and Fischer [17] 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. 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, most of these threats could have been prevented by timely and effective action to address the anger, pain, anxiety, or psychological impairment of perpetrators. Despite these compelling observations, no systematic methods or tools have yet been developed and validated to provide a predictive capability, although there are efforts underway (e.g., [5] [6] [10]). As noted in the introduction, challenges facing the security/counterintelligence community are not limited to the technical problems of model development, but also include obstacles to designing and conducting valid and robust evaluations of prospective tools. In this section, several key methodological issues and challenges are discussed.

A predictive system may be evaluated using historical data—the advantage is that ground truth and outcomes may be known. For many reasons, the use of historical data poses the most robust, and least controversial, strategy for testing predictive analytic methods. Whether detection or prediction are involved, there are ethical issues to be addressed (the reader is referred to [18] and [19] for thoughtful discussions of privacy and ethical issues underlying insider threat monitoring). But evaluation of a predictive system presents special challenges.

When the evaluation of a predictive system occurs in a real-time operational setting, there is the following dilemma: if the system identifies a high-risk individual or POI, but no abuse or crime has occurred, what (if any) action should be taken? The standard answer is that if no illegal acts or policy violation has occurred, then no action is called for. This may not be the best possible course, especially if we could apply some well-tested (mature) supporting technical tools to help the high-risk individuals without adversely impacting their rights. Yet, there are more questions and some pitfalls:

- Since the prediction is not perfect, the conclusion may be erroneous. Acting may harm the individual involved, and this may expose the organization to litigation.
- Confronting a POI with "evidence" of risk factors could be construed as harassment; this could exacerbate a possibly stressful situation and produce negative consequences.
- Legal and counterintelligence stakeholders might prefer that no action be taken, but that behavior/cyber monitoring should be increased, so as to collect "actionable" data that would hold up in court or to identify possible internal or external collaborators.
- On the other hand, cyber security and operational security stakeholders may prefer to take some sort of defensive action in order to protect and preserve human and organizational assets.

Beyond organizational pitfalls associated with acting on predictive analytic outputs, the application and testing of predictive methods pose difficult technical/scientific challenges because the very definition of "hit" and "false alarm" become clouded in this situation. Without ground truth data and definitive outcomes (identifying perpetrators), there is no basis for calculating P(hit) and P(false alarm). This difficulty makes the methods described in Section II *A* and *B* all the more worthy of consideration.

In light of these challenges in conducting evaluations of predictive systems, an assessment rubric might be considered that relies on expert judgment of risk (an example of this approach was described in Section II *A*). The rubric is discussed below:

- *Identify risk indicators.* In any scientific endeavor it is necessary to specify the dependent variables to be studied, and upon which the risk model is based. The indicators are not only identified, but also must be specified in sufficient detail to enable quantitative or qualitative data collection.
- *Develop risk metric*. The model must specify how the risk indicators shall be combined in the formulation of the risk model. As has been previously noted, several investigators have argued for development of a composite measure that combines risk scores from a variety of indicators [4] [5].
- *Estimate requirements for test population and test data collection.* Specify the test subjects to be included in the study. Will all individuals in the organization be participants in the study, or will the test be restricted to a subset of the population? If not all individuals are included, identify an unbiased method of selecting test subjects (e.g., random selection). Sampling biases undermine interpretation of results. Another question is: How many test subjects are required? Figure 5 provides a conceptual discussion and speculates on some answers.
- *Set up data collection protocols.* To support a proper scientific evaluation of the risk model, detailed plans must be described for collecting test data. Procedures must be developed to protect data integrity and privacy of individuals—especially with regard to de-identification of the data that requires converting personal ID numbers to coded identifiers, so that only a select group of organizational staff have the capability to "decode" the ID numbers for purposes of validating the predictions of the model.
- *Collect data.* Acquisition of data from varied sources (e.g., human resources/behavioral data, psychosocial data, cyber monitoring data, security records) occurs over the specified time interval selected for the study. Because the target activities are rare in terms of base rate as well as occurrences of abuse for a given individual, use of sufficiently long study duration is critical—e.g., minimum 12 months
- *Apply model to generate predictions*. Use obtained data as input to the model, to generate output. To prepare for validation against expert judgments, rank the obtained risk scores (with associated de-identified/coded participant IDs) generated by the model.
- *Obtain validation data for a test set.* To avoid bias, it is important to obtain expert judgments for a set of target individuals that includes not only the highest-risk outputs of the model, but also other representative members of the population (more detail on this validation step is provided below, in Section IV). For this set of individuals, obtain risk judgments from qualified experts within the organization (such as an inter-disciplinary team comprising human resources, management, security, cybersecurity, and legal representatives). Re-apply the de-identification process prior to returning this feedback to the model evaluation team.
- *Compare the expert risk judgments (rankings) with the rankings determined by the risk model.* The evaluation team may conduct various statistical analyses to assess the performance of the risk model. Some representative analyses were described in Section II *A* and *B*.

SAMPLE SIZE CONSIDERATIONS

We know that only a very small fraction of the population are POIs (say, 0.05% or even 0.01%).

- **Suppose we monitor 100% of a population of 100,000 individuals**
  - Assuming 0.05% are POIs, we expect to find 50 POIs:
    - An 80% uncertainty interval is 41-56 POIs
    - A 98% uncertainty interval is 34-67 POIs
  - Assuming 0.01% are POIs, we expect to find 10 POIs:
    - An 80% uncertainty interval is 6-14 POIs
    - A 98% uncertainty interval is 3-18 POIs
- **Suppose we sample 10% of the population; i.e., 10,000**
  - Assuming 0.05% are POIs, we expect to find 5 POIs:
    - P(sample will have 0 People of Interest) = 1%
    - P(sample will have ≤ 2 People of Interest) = 12%
    - P(sample will have ≤ 4 People of Interest) = 44%
  - Assuming 0.01% are POIs, we expect to find 1 POI:
    - P(sample will have 0 People of Interest) = 37%
    - P(sample will have ≤ 1 People of Interest) = 74%

Assuming a base rate of 0.01%, then sampling only 10,000 individuals out of a population of 100,000 yields a relatively high probability that we will not run across any POIs in our validation test. With such a low base rate, we are advised to sample at least 50,000 individuals out of a population of 100,000. In that case, P(sample will have ≤ 2 People of Interest) = 12%

Figure 5. Sample Size Considerations

IV. MEASURE OF EFFECTIVENESS

It has been noted that there are difficulties in obtaining performance metrics for various reasons. When appropriate data may be acquired to support outcome assessment based on actual cases (e.g., if historical data are available), measures of effectiveness may be derived using statistics such as associated with a confusion matrix and hit/false positive rates. In the absence of such data, and indeed to complement such analyses, an "impact assessment" methodology is advised. Using the evaluation rubric and test methodology described in the previous section, we now briefly consider measures of effectiveness. While an organization might comprise a total of $\boldsymbol{N}$ individuals, it might be that only a subset $\boldsymbol{N'}$ is analyzed by the algorithm (preferably $\boldsymbol{N'} = \boldsymbol{N}$). As described above, the proposed evaluation rubric generates an anonymized list of the $\boldsymbol{N_1}$ individuals that the model considers to be at the highest risk—to use less "charged" terminology, let us say that the model generates an anonymized list of "most interesting" individuals, those who might be considered most atypical of

the general population—we denote this as set ***A***. The test list will also include a random selection of N_2 anonymized individuals from the organization's general population (excluding set ***A***). There are at least two methods that could be used for evaluation:

- Method 1—all $N_1 + N_2$ individuals are listed in random order and the list is sent to the expert panel for evaluation. N_2 can be any size, preferably $N_2 \geq N_1$. The panel rates the extent to which each listed individual should be considered a POI. The analysis of the results should demonstrate if the use of the algorithm to generate set ***A*** yields a higher percentage of POIs identified by experts.
- Method 2—pairs of individuals are formed, one from the N_1 individuals in set ***A***, and one from a list of N_2 $(=N_1)$ individuals randomly selected from set $N'-A$. Each pair is presented to the experts, who must choose which of the two is more interesting. This is expected to be a much easier task for the experts.

We seek a measure of effectiveness based upon the expert feedback. Suppose that the expert evaluators determine that m_1 POIs out of the N_1 (in set ***A***) are actually of interest to them, and further assume that out of the other set of N_2 individuals, m_2 are deemed to be POIs by the organization—we denote these m_2 individuals as comprising set ***B***. As an illustrative example, suppose that $N' = 5000$, $N_1 = 50$, and $m_1 = 24$; and $N_2 = 50$ with $m_2 = 1$. Then the results are shown in Table I.

TABLE I. CONTINGENCY TABLE FOR AN ILLUSTRATIVE EXAMPLE

	B	~***B***		*Not Given to Panel*
A	$m_1 = 24$	$N_1 - m_1 = 26$	$N_1 = 50$	0
~***A***	$m_2 = 1$	$N_2 - m_2 = 49$	$N_2 = 50$	$N' - N_1 - N_2 = 4900$
	$m_1 + m_2 = 25$	$N_1+N_2 - m_1 - m_2 = 75$	$N_1+N_2 = 100$	

Note that the rightmost portion of the table (labeled "Not Given to Panel") reflects the subset of N' analyzed cases (out of a total of N in the population. The values of N' and N do not affect the metric calculations below, but they do impact the overall success in assessing the performance of an analytic tool, as discussed in Section III. It is also important to note that set ***A*** does not necessarily equal the set of all individuals identified in the sample of N' analyzed cases who are deemed POIs. Rather, set ***A*** is a ranked list based on computed POI risk. Stakeholders must set an "interest" threshold. In our example, a more stringent threshold would be to identify the top ten individuals out of the N_1 in set ***A***. A less stringent criterion would be to include the entire set of N_1 individuals in set ***A***. By setting varying thresholds between 1 and N_1, one may compute associated hit and false alarm probabilities to generate an ROC curve for the algorithm being tested. Generation of the ROC curve provides important evaluative insight into the utility of an algorithm.

Beyond this, a number of measures of effectiveness may be considered. Lenca et al. [20] described twenty alternative measures of association or rules applied to data mining algorithms. The idea behind the measures is the assertion that the greater the intersection of sets ***A*** and ***B***, and the fewer counter-examples (***A***,~***B***; ~***A***,***B***) to the rule there are, the better the rule. Some metrics considered in [20] are *support* = P(***A***,***B***); *lift* = P(***A***,***B***)/[P(***A***)P(***B***)]; and *Bayes Factor* (originally defined by Jeffreys [21]), *BF* = P(***B***|***A***)/P(***B***|~***A***). Note that *BF* is a nonzero real number, $0 \leq BF \leq +\infty$. Lenca et al. [20] used eight criteria to evaluate the measures, including normative properties of *asymmetry* and *independence* and subjective properties such as *intelligibility* (whether the rule's definition is easily understood and interpretable by domain experts). The *BF* metric was favorably ranked across all of the criteria.

We consider the *BF* measure to be quite interpretable and useful for assessing the utility of insider threat algorithms. At the very least, it provides a useful supplement to the ROC analysis. In our example, $BF = (m_1/N_1)/(m_2/N_2) = 24$. One might say that a tool with a *BF* score of 24 has a very strong impact in enriching the detection process beyond the current baseline; indeed, in this illustrative case, the algorithm being evaluated identified 24 times more POIs than a baseline method. Clearly this consideration addresses a property of prediction that is different from statistical significance. In this vein, we suggest an interpretation of the *BF* measure to reflect what we refer to as an *Enrichment Ratio (ER)*. Specifically, $ER = BF = (m_1/N_1)/(m_2/N_2)$. The value of 24 obtained in the example means that the unaided method would have to examine 24 *times* the number of individuals than would be examined by the method being evaluated in order to expect to find the same number of POIs. This reveals a practical impact on labor hours. Clearly all organizations have a nonzero risk of insider threats; large organizations may go to great expense to find insider threats. Suppose that an organization dedicates ten staff members (full-time equivalents, FTEs) to this effort. As shown in the example of Table II, if the *ER* for an insider threat risk model or tool is 2:1, then use of the tool might be expected to enable the organization to find the same number of insiders with just five FTEs. This saves five FTEs, or $1.25 Million (assuming $250K/FTE).

TABLE II. ENRICHMENT RATIO: PRACTICAL INTERPRETATION FOR IMPACT ASSESSMENT OF A PROPOSED TOOL

BF or Enrichment Ratio Score	Anticipated number of FTEs to get approximately equal number of "finds"	Savings (assuming $250K/FTE) compared to using 10 FTEs	Anticipated increase in individuals "found" using the proposed tool table
1:1	10	0 FTEs = $0K	0
2:1	5	5 FTEs = $1,250K	Double
5:1	2	8 FTEs = $2,000K	Quintuple
10:1	1	9 FTEs = $2,250K	Ten-fold Increase

V. CONCLUSIONS

The insider threat is a prime security concern for government and industry organizations. Despite much research focused on insider threat risk models and tools to detect or mitigate insider attacks, development and validation of tools still ranks among the most critical research needs. As insider threat programs come into operational practice, there is a continuing need to assess the effectiveness of tools, methods, and data sources, which enables continual process improvement. Best practices demand that analytic processes be measured for their effectiveness, preferably by calculating both true positives and false positives over a specified period of time [3]. This is particularly challenging in operational environments, where the actual number of malicious insiders in a study sample is not known. In the present paper, we have attempted to address the difficult challenge of developing evaluation strategies and measures of effectiveness.

Insider threat programs best practices also stress the importance of managing the time of analytical/threat analysis staff, which typically deals mostly with activities on the low end of the complexity and damage spectrums [3]; programs should structure their metrics to help distinguish effective vs. ineffective detectors by identifying detectors with too few analytical outcomes [3]. In this vein, we briefly discussed measures of effectiveness, including a *BF* measure [21] used prominently in data mining research [20] that supports our application of an *Enrichment Ratio (ER)* to assess the practical impact of proposed tools. The *BF* measure with the *ER* interpretation quantifies the savings in terms of staff hours or labor/analysis costs that may be attributed to the application of a threat detection or mitigation tool. By comparing the amount of data that must be monitored by a baseline program in order to achieve comparable performance to a proposed tool, this measure assesses the proposed solution in concrete operational terms. It is hoped that this discussion of evaluation methodology, analytic approaches, and metrics will help to advance the progress of insider threat mitigation research.

ACKNOWLEDGMENT

The authors thank an anonymous reviewer for suggesting linkages with metrics used in data mining research.

REFERENCES

[1] D. Capelli, A. Moore and R. Trzeciak, *The CERT Guide to Insider Threats,* Upper Saddle River, NJ: Addison-Wesley, 2012.

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

[3] M. Guido and M. Brooks, "Insider threat program best practices," in *46th Hawaii International Conference on System Sciences*, Maui, Hawaii, 2013.

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

[5] F. Greitzer and D. Frincke, "Combining traditional computer security audit data with psychosocial data: predictive modeling for insider threat," in *Insider Threats in Cyber Security*, New York, NY: Springer, 2010, pp. 85-114.

[6] F. Greitzer, L. Kangas, C. Noonan, A. Dalton and R. Hohimer, "Identifying at-risk employees: a behavioral model for predicting potential insider threats.," in *45th Hawaii International Conference on System Sciences (HICSS-45)*, Wailea, Maui, Hawaii, 2012.

[7] C. Langin and S. Rahimi, "Soft computing in intrusion dection: the state of the art," *Journal of Ambient Intelligence and Humanized Computing,* vol. 1, pp. 133-145, 2010.

[8] K. Ilgun, R. Kemmerer and P. Porras, "State transition analysis: a rule-based intrusion detection approach," *IEEE Transactions on Software Engineering, vol. 21*, pp. 181-199, 1995.

[9] F. Greitzer, L. Kangas, C. F. Noonan, C. R. Brown and T. Ferryman, "Psychosocial modeling of insider threat risk based on behavioral and word use analysis," *e-Services Journal*, 2013 (in press).

[10] C. Brown, A. Watkins and F. Greitzer, "Predicting insider threat risks through linguistic analysis of electronic communication.," in *46th Hawaii International Conference on Systems Sciences (HICSS-46)*, Wailea, Maui, Hawaii, 2013.

[11] Y. R. Tausczik and J. W. Pennebaker, "The Psychological Meaning of Words: LIWC and Computerized Text Analysis Methods," *Journal of Language and Social Psychology, vol. 29*, no. 1, p. 24054, 2010.

[12] D. Vysochanskij and Y. Petunin, "Justification of the 3σ rule for unimodal distributions.," *Theory of Probability and Mathematical Statistics, vol. 21*, pp. 25-36, 1980.

[13] P. Mahalanobis, "On the generalised distance in statistics," *Proceedings of the National Institute of Sciences of India, vol. 2*, no. 1, pp. 49-55, 1936.

[14] D. Green and J. Swets, *Signal detection theory and psychophysics*, New York: Wiley, 1966.

[15] I. Pollack and L. and Decker, "Confidence ratings, message reception, and the receiver operating characteristic," *Journal of the Acoustical Society of America, vol. 31*, pp. 1500-1508, 1958.

[16] T. Wickens, *Elementary signal detection theory*, New York: Oxford University Press, 2002.

[17] E. D. Shaw and L. F. Fischer, "Ten Tales of Betrayal: The Threat to Corporate Infrastructures by Information Technology Insiders Analysis and Observations," Defense Personnel Security Research Center, Monterey, CA, 2005.

[18] A. Kiser, T. Porter and D. Vequist, "Employee monitoring and ethics: Can they co-exist?," *International Journal of Digital Literacy and Digital Competence, vol.* 1, no. 3, pp. 30-45, 2010.

[19] F. L. Greitzer, D. A. Frincke and M. Zabriskie, "Social / Ethical Issues in Predictive Insider Threat Monitoring," in *Information Assurance and Security Ethics in Complex Systems: Interdisciplinary Perspectives,* Hershey - New York, Information Science Reference, 2011, pp. 132-161.

[20] P. Lenca, P. Meyer, B. Vaillant, and S. Lallich. "On selecting interestingness measures for association rules: user oriented description and multiple criteria decision aid." *European Journal of Operational Research, 184 (2)*, pp. 610–626, 2008.

[21] H. Jeffreys. "Some tests of significance, treated by the theory of probability." *Proceedings of the Cambridge Philosophy Society, 31*, 203-222, 1935.