Supervised and Unsupervised methods to detect Insider Threat from Enterprise Social and Online Activity Data

Gaurang Gavai[1]*, Kumar Sricharan[1], Dave Gunning[1], John Hanley[1], Mudita Singhal[1], and Rob Rolleston[2]
[1]*Palo Alto Research Center, Palo Alto CA 94304 USA*
{ggavai, skumar, dgunning, jhanley, msinghal}@parc.com
[2]*Palo Alto Research Center East, Webster NY 14580 USA*
rrolleston@parc.com

Abstract

Insider threat is a significant security risk for organizations, and detection of insider threat is of paramount concern to organizations. In this paper, we attempt to discover insider threat by analyzing enterprise social and online activity data of employees. To this end, we process and extract relevant features that are possibly indicative of insider threat behavior. This includes features extracted from social data including email communication patterns and content, and online activity data such as web browsing patterns, email frequency, and file and machine access patterns. Subsequently, we take two approaches to detect insider threat: (i) an unsupervised approach where we identify statistically abnormal behavior with respect to these features using state-of-the-art anomaly detection methods, and (ii) a supervised approach where we use labels indicating when employees quit the company as a proxy for insider threat activity to design a classifier. We test our approach on a real world data set with artificially injected insider threat events. We obtain a ROC score of 0.77 for the unsupervised approach, and a classification accuracy of 73.4% for the supervised approach. These results indicate that our proposed approaches are fairly successful in identifying insider threat events. Finally, we build a visualization dashboard that enables managers and HR personnel to quickly identify employees with high threat risk scores which will enable them to take suitable preventive measures and limit security risk.

Keywords: Anomaly detection, insider threat detection, quitting detection, enterprise social data

1 Introduction

Insider threats are threats with malicious intent directed towards organizations by people internal to the organization. These include physical sabotage activities, theft of confidential data and business secrets, and fraud. Insider threat activities pose a severe challenge to the well-being of an organization, and it is naturally critical for organizations to guard against such events.

For these reasons the insider threat detection problem has been studied extensively [1]. This includes work done in the cyber-security realm, social science, business management and other relevant areas. Recently, we have seen the advent of data analytics based methods for identifying insider threat events. Compared to the traditional approaches of post-attack analysis and subsequent change of policy, predicting threats from data offers the benefit of continuous and in-time evaluation. It circumvents the various issues plaguing the traditional approach such as their significant administrative overhead and their infrequency. The questions here are rather, how strongly a real-world insider threat event correlates with enterprise data traces, and whether the prediction is reliable enough to serve the purpose of risk management.

Journal of Wireless Mobile Networks, Ubiquitous Computing, and Dependable Applications, volume: 6, number: 4, pp. 47-63
*Corresponding author: Tel: +1-(650)-812-4760

To this end, we have developed a privacy preserving feature extraction approach to capture a rich set of online behaviors that can be indicative of insider threat behavior, with each of these features being free of personally identifiable information.[2] Next, we take two approaches to detecting insider threat: (i) we have applied a state-of-the-art unsupervised anomaly detection method called isolation forest [3] to identify anomalous events and (ii) we have developed a classifier that uses labels of employees who have quit the organization as a proxy for potential insider threat activity. The use of quitting labels as proxy for insider threat activity is reasonable in light of analysis that quitting prediction may also inform risk management. For instance, Infosecurity Magazine [4] reports that "50% of job leavers are likely to steal confidential company data".

Unsupervised methods for insider threat detection include the work done by Mathew et al. [5] based on user access patterns, the work of Eberle et al. [6] on using social graphs, and several papers based on behavioral models [7, 8, 9]. Hoda et al. [10] detect peer groups of users and modeling user behavior with respect to these peer groups, and subsequently detect insider activity by identifying users who deviate from their peers with respect to the user behavior models. There have also been various approaches based on detection of specific insider threat scenarios [11, 12]. Our approach is based on design of novel features from enterprise data that are reflective of insider threat behavior, and subsequent analysis of this data to identify inconsistent, statistically rare behavior that can be indicative of insider threat activity [2]. Our method differs from the work of Eberle et al. in that they focus only on social graphs, while our features are more inclusive. Similarly, Mathew et al. focus only on user access patterns. Also, the work by [11, 12] focuses on specific scenarios only, and as a result, they will be unable to detect previously unknown insider threat scenarios. In contrast, our method can detect previously unknown types of insider threat activity because our method is based on identifying statistically abnormal behavior and is not reliant on having prior knowledge of insider threat scenarios. Finally, our approach differs from the work by [7, 8, 9, 10] in that our approach does not require the estimation of normal behavior models. Rather, our method directly identifies statistically abnormal behavior. A closely related body of work to our paper is the work by Young et al. [13], who also employ anomaly detection methods to detect insider activity. Our work differs from theirs in that they determine anomalies only by contrasting user attributes against their peers at every time instance, while we also detect insider activity by identifying abnormal changes in user attributes over time. This helps us to also identify users whose behavior might appear normal at every time instance when looked at individually but exhibit anomalous trends across time, possibly indicating insider activity. This concept of across-time consistency is describe in detail in the paper by Hoda et al. [14].

On the other hand, for the supervised approach, the employee churn problem has been studied extensively by business management, economists, psychologists, and social scientists. Ample broad findings can be found in the literature. For instance, company size, industry and pay scales play a key role in determining attrition rate. Industries that largely employ unskilled labor have a higher rate of attrition as compared to those that largely require skilled labor. Attrition rate is also highest amongst the lowest paying jobs. While these findings provide many qualitative insights, they have not yet reached the level of mathematical predictors that can be deployed to perform churn analysis and prediction. Our earlier work [15] has made a first attempt at using a data-driven approach to study the problem of predicting attrition within an organization. We have constructed models to predict if and when an employee is likely to quit the company using email features. In this paper we extend the analysis to a broader set of data: posts, messages, and group conversations within a much more comprehensive social space. Another piece of related work is described in [16]. The authors use LinkedIn data to provide job-switching recommendations to users. Their work models a user's tenure in their current employment with a proportional hazard model, and uses it to decide when to provide job recommendations. In a sense this is the opposite of what we are trying to do.

Pros and Cons between supervised and unsupervised approaches	
Supervised approach	Unsupervised approach
(p) Higher precision rate (wrt quitting)	(p) Higher recall rate
(c) Optimized wrt quitting instead of insider threat	(c) Unsupervised approach detects all abnormal behavior

Table 1: Table listing pros and cons of the two methods. (p) indicates pros and (c) indicates cons

Next, we contrast and compare the supervised and unsupervised approaches. For the unsupervised approach, given that we are employing statistical anomaly detection methods on a broad set of features, not all anomalies that are detected will necessarily be related to insider threat activity. However, the broad feature selection ensures that our recall rate will remain high, at the cost of lower precision. On the flip side, the supervised method has higher precision, but with respect to the quitting behavior which is only a proxy for possible insider threat activity. The two methods therefore complement each other. The pros and cons of these methods are summarized in Table 1.

We test our dual-pronged approach on a real world data set named 'Vegas'. The Vegas data set is a proprietary collection of social and activity data traces such as email, work practice logs, records of file access, device usage, and web browsing patterns for all employees belonging to a sub-division of a large corporate organization. Social communications between the employees such as email and instant messages are an important part of this data set. In order to use this data set for evaluating insider threat detection capabilities, insider threat activity behavior is artificially injected into the data via a collection of red team users [2]. Details about this artificial injection process are given in Section 2 and the Appendix.

We obtain a ROC score of 0.77 for the unsupervised anomaly detection approach and an accuracy score of 73.4% using the supervised approach on Vegas. Given the difficulty of the insider threat detection problem, these results are fairly encouraging. Finally, we build a visualization dashboard that charts the predicted anomaly scores for all the employees in an organization. This dashboard makes it easy for end-user analysts to quickly identify users with high anomaly risk in order to take suitable preventive measures.

The rest of this paper is organized as follows. We describe data cleaning and pre-processing steps and our privacy preserving feature extraction approach in Section 2. The unsupervised anomaly detection method, the supervised classification method, and the corresponding results are described in Sections 3, 4, and 5, respectively. We describe our visualization dashboard in Section 5 and finally, give our conclusions in Section 6.

2 Feature extraction and data

We ran our experiments on a real-world dataset which we refer to from here on as Vegas. It was specifically collected in order to evaluate detection methods for insider threat activity[1]. This multi-domain employee dataset containing many modalities of employee information (emails, application logs, login information, business unit hierarchy, etc.) was sourced from a single business unit of a larger organization. It is measured over 10 months, from October 2013 through July 2014, and comprised of 6805 distinct users with over a billion user activity records.

[1] http://www.networkworld.com/article/2220363/security/darpa-expands-insider-threat-research.html

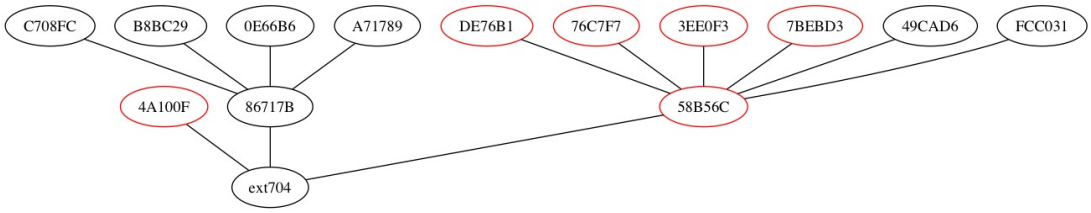

Figure 1: Example Sub-tree in Hierarchy. Each node represents a user and red nodes are quitters.

The CERT organization from SEI was tasked with the job of intelligently inserting synthetically created malicious events corresponding to a subset of "red team" users into the data. The team employed various complicated models to make the fictional scenarios mirror real world ones as closely as possible. [17]. Additionally, to ensure that algorithms were being tuned to find anomalous data and not just artificial data, benign events were injected as part of user histories as well. The full list of scenarios and their descriptions can be found in Appendix A of the paper. The aim of unsupervised methods in this paper is to find these Red Team users.

For our supervised approach, we used the employees who quit the company as a proxy for insider threat [18]. We framed the problem formally as: For a given user U and a day index T, predict if that user U *will quit company at some time instance* T by studying the user's activity from T_{begin} up to T [18]. We were provided daily hierarchy snapshots of the business by the company and we used the quitting labels from these to identify the quitters. Additionally there were a few users whose IDs were removed without explanation from the hierarchy snapshots with all records of their various activities ceasing to exist from that point. We treated these users as quitters too but labeled them differently as *pseudo-quitters* since we have no formal record of them quitting. We identified 555 quitters and 1270 pseudo-quitters in the data.

2.1 Pre-processing

The organization that provided us the data from it's business unit had several privacy concerns due to it's sensitive nature. To alleviate these, we created a security protocol in conjunction with the organization to ensure that the required level of anonymization was maintained across all our experiments. This however did lead to some reconstruction issues [2].

We also needed to build a hierarchy of the employees within the business unit in conjunction with the features that we wanted to create. To achieve this, we used the snapshots of daily hierarchies that we had access to and constructed a normalized version over time in which the most persistent relationships between employees and their direct supervisors were maintained. Unfortunately, since the business unit we were studying did not include any of the highest levels of the organization, this resulted in over 200 distinct, disconnected sub-trees as opposed to a single structured tree. We selected a few, coherent ones from these for comparative and visualization purposes. An example of these can be seen in Figure 1.

2.2 Features

An employee's electronic presence, or at least it's most important aspects, can be encapsulated in their email, web, login and application usage behavior. Consequently, this implies that most insider threat activity or anomalous behavior with an electronic footprint can be detected from features constructed from employee traces in these information domains.

To aid our feature design process, we also conducted interviews with recent job quitters [18]. We interview 12 people with varied job roles and gained various insights which varied from easily intuitive

Table 2: Vegas features.

Email Usage Features
Emails Sent/Read Weekly Count
No. of Emails sent at Daytime
No. of Emails sent at Night
No. of Emails read at Daytime
No. of Emails read at Night
Email Content Features
No. of exclamation marks, question marks, multiple marks, ellipses, commas, semi-colons dashes, double dashes, brackets, colons
Avg Word Length in Subject
Avg Character Length in Subject
Avg Word Length in Content
Avg Character Length in Content
Log-on log-off Features
No. of log-ons
No. of log-offs
No. of hours with log-on activity
No. of hours with log-off activity
Activity Features
No. of activity types
Max contiguous hours spent on activity
Number of activities
Time spent on Email apps
Time spent on Productivity apps
Time spent on Web apps
Time spent on Engineering apps
Web Usage Features
Total time spent on websites
Time spent on career sites, web mail, entertainment sites, internal social media, internal sites, news sites, search sites, private social media, tech sites

to considerably counter-intuitive. This drove the formulation of our email content feature set [18]. Some of the insights we gained were:

- Increased use of career sites.
- Increased use of personal email (for job applications).
- Decreased attention span at work resulting in more constant task switching.
- Efforts are often made by quitters to maintain normal behavior at work resulting in a more neutral sentiment in emails.
- Shortened work hours while seeking a new job to make time for interviews, but regular hours in the last few weeks once a new job was found.

Using the afore-mentioned insights, we engineered a total of 42 features over five domains or categories: email usage, email content, log-on log-off behavior, application activity and web activity. We explicitly note that this feature set is general and applicable to datasets and scenarios other than Vegas. The full set of features are listed in Table 2. We describe each of these five categories in further detail below.

We calculate features for each user U with respect to each individual day indexed by T. Thus for every (U,T) tuple here, the corresponding value of each feature was on a daily basis. There were a number of motivations for doing so including allowing the dataset to evince activity between and outside set work hours and contrast the difference between different types of days (weekdays/weekends/holidays).

(i) Email Usage Features: The email usage features describe how users manage communicating to people within and outside the business unit using their business email portal. They are aggregated on a daily basis and can be used to capture evidence of abnormal collusion, since many Red Teams events hinge on the same. An unnatural increase in the flow or volume of email either in terms of time or people could be indicative of the same. Example Scenarios: Insider Startup, Indecent RFP

(ii) Email Content Features: The email content features are created from the body of the emails that were sent by business unit users to people both within and outside the unit. The email usage features section covers the reasons one would want to capture abnormal collusion by business unit users. Content-based methods such as sentiment analysis and speech act scores would greatly aid in being able to discern the modality of this abnormal communication and how much it deviates from the norm. There are some daily aggregated count features for punctuation marks etc. that are calculated in order to be able to capture changes in writing style that may occur as a result of users attempting to cover their malicious tracks. Their inclusion is further justified in this scenario since recall is valued much higher over precision. Example Scenarios: Indecent RFP, Czech Mate

(iii) Web Browsing Usage Features: The web browsing usage features describe the categorized browsing histories of Vegas users. We constructed categories of websites and ascribed popular variants of each kind to the category. Then we calculated the daily aggregated time spent by individual employees in each category. For example, time spent on career sites, email sites, news sites, social media sites, etc. Time spent on non-job related sites could be a useful barometer to the level of disengagement of employees and the flux therein can capture the variance in the degree of satisfaction they derive from their jobs. More dissatisfied employees have a greater likelihood of being an insider threat [4]. Example Scenarios: Naughty By Proxy, Hiding Undue Affluence

(iv) Application Activity Features: The application activity features log the usage of different categories of applications by employees. Application codes were divided into four classes: Email, Web, Productivity, and Engineering. Email and Web refer to application clients used for the same whereas Productivity includes note-taking, document manipulation, calculators etc and Engineering revolves around coding/ programming tools. We aggregated the time spent in each category on a daily basis. This category was created to capture insider threat scenarios that involve employees using permitted applications for excessive periods of time that may indicate them committing actions outside their job duties. Example Scenarios: Masquerading, Masquerading 2

(v) Employee Log-on/Log-off Features: The log-on/log-off features are derived from the sign-ins of employees. They are aggregated on a daily basis as the number of hours with such activity and the total number of log-ons and log-offs. Insiders are more likely to commit their activities at odd times when there are less people physically present so as to avoid being detected or perhaps conversely suddenly become extremely punctual to avoid detection. This subset of features was created to evince such flux either with respect to the employee's own past history or with respect to their peers in the organization. Example Scenarios: Stealing Login Credentials, Exfiltration Prior to Termination

We cast a wider net of features than those that can be directly correlated to quitting behavior or insider threat activity, because the aim of our analyses is a high recall rate, regardless of it being at the cost of

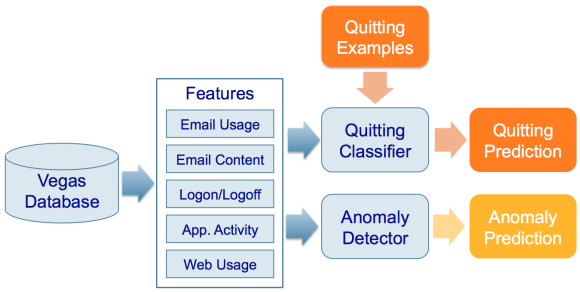

Figure 2: Framework for identifying anomalous users and quitters.

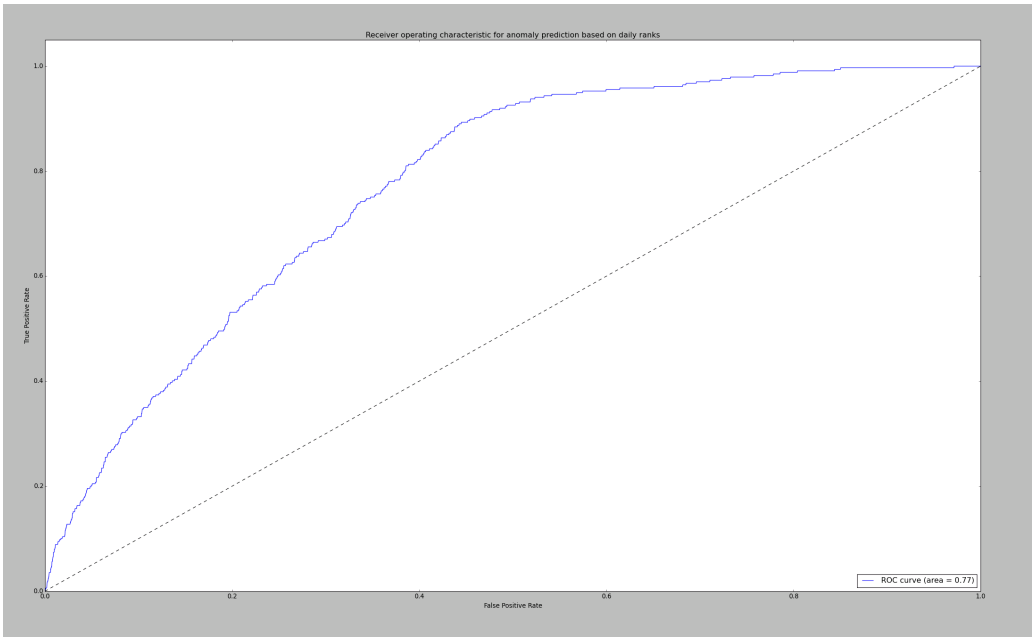

Figure 3: Quitter Prediction ROC curve for Vegas. x-axis is 'false positive rate', y-axis is the 'true positive rate'. The area under the ROC curve is 0.77.

a loss in precision. At the same time, the list of features is non-exhaustive due to numerous anonymity and privacy related constraints imposed on us by the organization owning the data.

In the next two sections, we describe our unsupervised and supervised approaches that exploit these features to detect insider threat. The framework of our analyses to predict insider threat is illustrated in Figure 2.

3 Unsupervised Approach: Insider Threat Detection

In this section, we describe our analyses in identifying users who likely pose an insider threat to the organization. We do so using the set of features that were identified and constructed in the previous section. Since it is exceedingly difficult in real-world insider threat scenarios to obtain ground truth for such activity, we pose this prediction problem as an unsupervised one. We state the formal problem statement of this experiment as: "At any given time instance, given their past online activity, predict if an employee is behaving abnormally either with respect to to his past activity, or with respect to the behavior of his peers." [2]

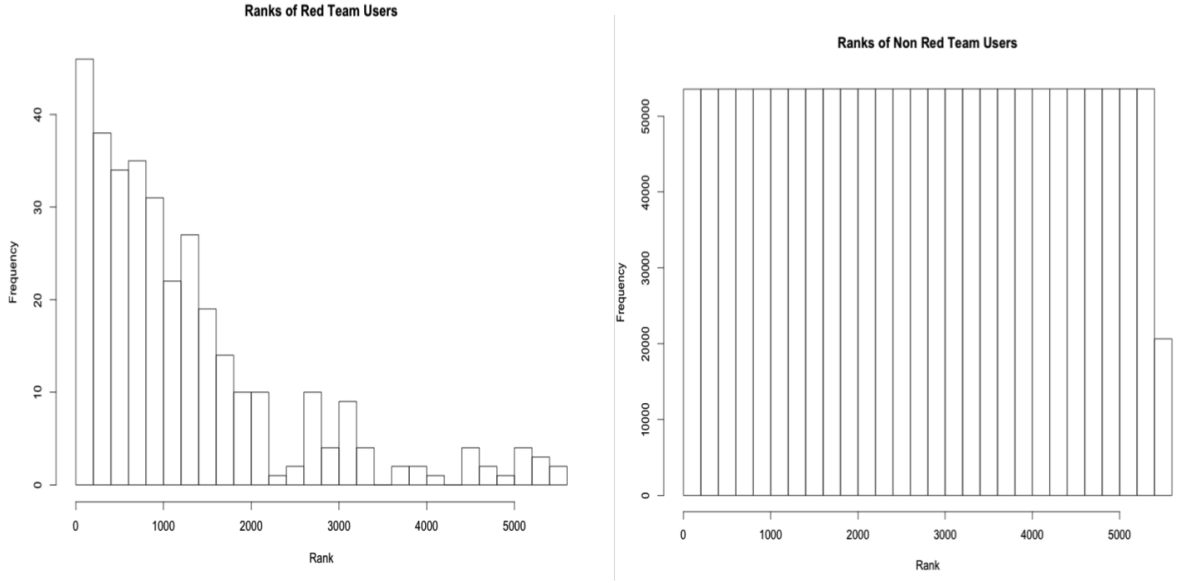

Figure 4: Comparison of histograms of daily rank using iForest method. Observe that the red team events on average have a higher rank (lower numerical value is considered to be higher rank e.g. 2 is a higher rank than 5), i.e., iForest assigns a majority of the red team events a high rank.

3.1 Feature manipulation

As mentioned previously, the Vegas dataset encompasses $n_U = 6805$ users from October 2013 to July 2014 (a period of 10 months) and can be specified as covering $n_T = 302$ days. Additionally, the set of features we constructed in the previous section comprises $n_K = 42$ distinct features. Thus, we can represent our entire feature set F as a tensor with dimensions as $n_U \times n_K \times n_T$.

Our aim is to use F and determine two distinct anomalous behaviors in the data set. Namely:

1. behavior of a user that is anomalous with respect to their peers

2. behavior of a user that is anomalous with respect to their own past activity

In order to do this, we used the feature set F to construct a data set D. We did this by populating the data set D with n_T matrices $M(T)$ for $T = 1, \ldots, n_T$. Every $M(T)$ matrix is of the dimensions $n_U \times 3n_K$. We construct these $M(T)$ matrices using the equations enumerated below:

1. Average:
$$M_1(T) = \frac{\sum_{i=1}^{T} F[:,:,i]}{T}.$$

2. Weighted average:
$$M_2(T) = \frac{\sum_{i=1}^{T} F[:,:,i] \exp\left[(i-T)^2/\sigma^2\right]}{T}.$$

3. Weighted difference:
$$M_3(T) = \frac{\sum_{i=1}^{T} (F[:,:,i] - F[:,:,i-1]) \exp\left[(i-T)^2/\sigma^2\right]}{T},$$

where $F[:,:,0]$ is set to 0, and σ is a weight parameter that controls the influence of past time instances. Here, we use $\sigma = 1$. We can easily observe that the $M_1(T), M_2(T)$ and $M_3(T)$ matrices are all of the same dimensions namely, $n_U \times n_K$. The intuition here is that the matrices $M_1(T)$ and $M_2(T)$ are effective in contrasting every user U with respect to their peers in the organization on the basis of the average of all past behavior and only recent past behavior with respect to each time instance respectively. Complementarily, the matrix $M_3(T)$ is useful in capturing anomalous behavior of every user with regards to the average of their own past behavior.

We denote $M(T) = [M_1(T), M_2(T), M_3(T)]$ for each $T = 1, .., n_T$. Thus, one can see that considering every $M(T)$, and applying the anomaly detection algorithm for the same, will result in us being able to identify both previously declared modalities of a user's anomalous behavior.

3.2 Anomaly detection algorithm

The algorithm used in the previous sub-section to detect the two types of anomalies in each time-step using the three matrix variants of the features is a modified version of the isolation forest algorithm [3], which is the current state-of-the-art in anomaly detection. We describe the algorithm in more detail here. This algorithm can be used to detect statistical anomalies in data. It does so by constructing a forest of isolation trees which are basically binary trees. These trees grow by recursively partitioning the data by selecting a threshold for a randomly picked feature from the set at every node. This process continues till each individual point in the data set is isolated in a leaf. Every point is ascribed a score that relates to the average number of splits required over the forest of isolation trees to isolate it into a leaf node.

Our modification to this algorithm is in that we not only identify the anomalous points that arise from the averaging across the forest, but also record the corresponding features that are the reason for their isolation. Our method to do so is to aggregate the extent to which every feature in every dimension of every tree isolates a point. We use this modified isolation forest and apply it to the previously constructed $M(T)$ sets at every time instance to generate a ranking of users at said instance. Due to our modification, we also have a motivation of why each anomalous user was flagged in terms of the feature that indicated so. We obtain this daily ranking by sorting in descending order the anomaly scores that the isolation forest generates from the $M(T)$ sets.

3.3 Evaluation

In this unsupervised approach, we only use the labels from the red team ground truth to evaluate the results of our modified isolation forest algorithm. To do so, we compare the red team member's daily ranks at every time instance to the non-red team members'. The histograms from this comparison are shown in Figure 4. We can clearly evince from these histograms that our modified isolation forest method is indeed successful in the identification of a majority of the anomalous insider threats since it assigns them a higher daily rank. Since the algorithm assigns a majority of the target events from the red team population high ranks, it is indicative of the fact that it can effectively identify insider threat activity on a conceptual level. Conversely, if the algorithm was not able to do so, the resulting histograms for both populations would have looked like the histogram of non-red team users does now. That is, the ranks assigned across the population would have been identically uniform. The most informative features, on the basis of the modified feature descriptor in the algorithm, were found to be the features derived from the content of employee emails.

The resulting ROC curve from the daily ranks of users (with an encouragingly high area under the ROC curve of 0.77 given the undefined, variegated nature of the anomalies we are trying to find) generated by the modified isolation forest algorithm is shown in Figure 5.

4 Supervised Approach: Quitting Detection

In this section, we describe our quitting analyses on the Vegas dataset using those instances in our data where user U quits in time T, $T+1$ or $T+2$ as the quitting label of the positive class. [18]

Out of the 6805 users in Vegas that represent a unit of the large organization we are studying, there are totally $555+1270=1825$ users + pseudo-quitters that quit at some point in the 10 months of data we possess. Approximately, this results in around $\sim 6K$ instances of the positive class, and $\sim 1M$ (U,T) instances of the negative class.

4.1 Feature importance

Primarily, we needed to identify the subset of features from our exhaustive set that were the most informative or important. To that end, we computed the mutual information (MI) [19] of the target positive class labels and each feature individually. The number of bits on average conveyed by an individual feature with regards to the target label is known as mutual information. The results of this information analysis indicated that the set of email content features or features that were derived from the content body of employee emails were the most informative with regards to the task of predicting quitters.

4.2 Classification

We set up the Vegas quitter prediction problem in a time-ordered fashion, The training set used to build a model for the prediction task was the first three months from October to December of 2013. The test data set was from January to July 2014. The model built from the training data was applied to this test set. We trained a number of classifiers on the training set and from that collection the best performing classifier relative to the set were random forests [20]. We used the random forests algorithm and created a model that was then applied to the test set and ran many iterations of the procedure so that we may account for the randomness existing in sampling. The classifier's resulting confusion matrix is shown in Table 3. The resulting normalized accuracy was 73.4%, and the recall rate was 72%. The ROC curve that corresponds to this experiment with an area under the curve of 0.76 is shown in Figure 5.

Table 3: Normalized confusion matrix for company quitter prediction on imbalanced data in Vegas.

		Prediction	
	Class	Normal	Quit
Truth	Normal	37.3	12.7
	Quit	13.9	36.1

5 Visualization

The analyses we presented in the previous section outputs a threat score for the anomaly detection piece and form the quitting detection piece, the likelihood of quitting for every user U at every time instance T . However as mentioned before, these scores are not important independently, but in comparison with the user's own history and that of their peers. For example, if an entire team is working to fulfill a deadline, all their scores would indicate a high level of anomalousness. Similarly, an employee with a non-traditional job role may show high anomalousness consistently. To make it easier to compare

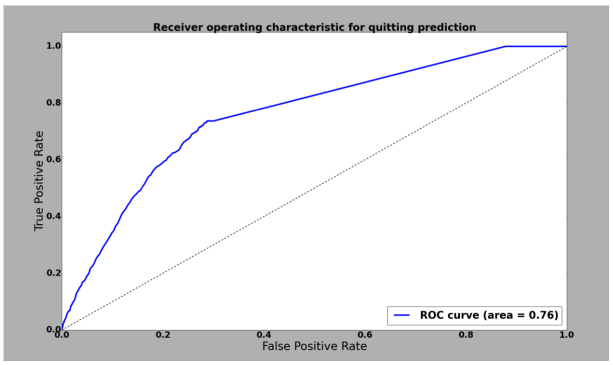

Figure 5: Quitter prediction ROC for Vegas.

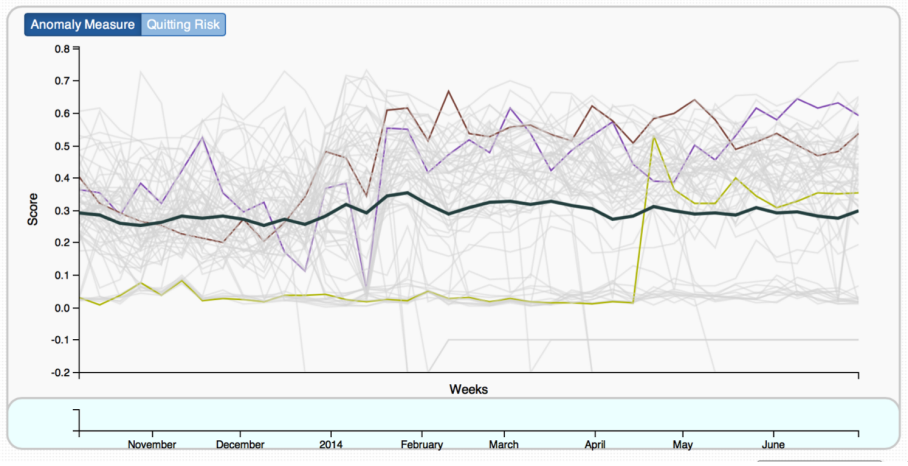

Figure 6: Visualization for identifying insider threat. Threat scores of three known insider threat users are shown in brown, purple and yellow, along with the average threat scores for their co-workers (black line).

the threat scores of users in order to identify those which an analyst should focus on, we have built a visualization tool that helps identify users who have abnormally large relative anomaly scores.

Recall that we can identify within division co-workers from the hierarchy structure we recovered. We built a visualization tool where users are organized with respect to this hierarchy. Furthermore, upon selecting any user of interest, his threat or quitting scores across time are highlighted along with the average threat or quitting score of his co-workers within the same division, allowing for quick identification of users who have relatively high threat scores compared to their peers [2].

Figure 6 shows a screenshot of our tool. In this screen-shot, we show the threat scores of two users along with the average threat scores for their co-workers (black line).

The aim of this prototype was to allow managers to be able to monitor the comparative state of their employees. However, it can additionally be used in an exploratory fashion to discover trends in quitting or anomalous behavior across the unit, organization or even in general. During our exploratory analyses we came a cross a trend of a ”double-crest” structure in the quitting score curves for all quitters. One can evince this in Figure 7, where we plot the quitting scores for 4 different quitters in the Vegas data set as a function of time. The interesting thing to note is that this ties in perfectly with the following observation that was made from the survey gathered from our conducted interviews. Namely that, in the job seeking phase quitters would show high quitting scores due to irregular activities, then a lull when they tried to

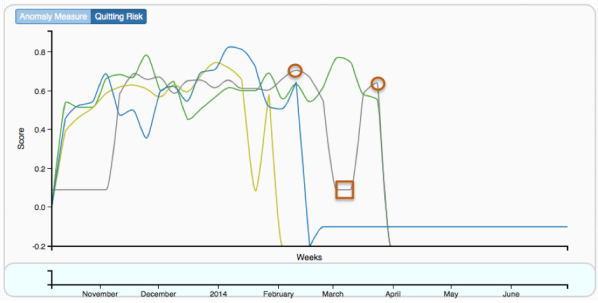

Figure 7: Quitting score profiles for 4 quitters in the Vegas data set. Double-crest structure can be observed. The two crests and the trough for the quitting score profile color coded in gray is highlighted by the two orange circles and the orange square respectively.

maintain a semblance of normalcy once that phase is over, followed by a spike again due to activities required to be conducted just prior to leaving an organization.[18]

6 Discussion and Future Work

In this paper, we have developed a data driven approach to detecting insider threat activity. First, we focused on engineering a set of novel features designed to capture behavior indicative of insider threat. The features we have developed include those based on email content, work-practice and online activity. We then used these features to detect insider threat using a two-pronged approach: (i) we used these features as input to an unsupervised anomaly detection method in order to detect suspicious behavior and (ii) we used these features in conjunction with quitting labels to develop a classifier using supervised classification methods.

From our experiments, we find that these approaches works well in detecting insider threat activity and quitting activity respectively. We are currently in the process of evaluating the precision of the quitting classifier with respect to insider threat activity instead of quitting activity. We expect that the precision will go down because quitting is only a proxy for insider threat behavior, but are hopeful that the quitting and insider events are correlated strongly enough for the precision of the quitting classifier to remain fairly high for detecting insider threat events.

Our results using these two methods are promising, and can be used to take preemptive steps to prevent people who have been identified by our models from engaging in insider threat / quitting activity. To this end, we have developed a dashboard front-end to enable managers and HR personnel to identify employees who are exhibiting a high risk of insider threat/quitting activity, and allow for suitable remedial measures to be taken. A possible issue with the current visualization scheme is that it works well when a small set of users are being explored, but can become cluttered when the number of users being investigated becomes large. We are looking at means to rectify this flaw by adding textual output capabilities to our tool so as to provide users with the flexibility of switching between the visual display and text-based drill-down of results. We intend to use ethnographic methods [21] to guide the addition of text output capabilities and understand the effectiveness of our system.

Acknowledgment

This research is funded in part by DARPA/ADAMS program under contract W911NF-11-C-0216. 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] T. F. Lunt, "A survey of intrusion detection techniques," *Computers & Security*, vol. 12, no. 4, pp. 405–418, 1993.

[2] G. Gavai, K. Sricharan, D. Gunning, R. Rolleston, M. Singhal, J. Hanley, and R. Rolleston, "Detecting insider threat from enterprise social and online activity data," in *Proc. of the 7th ACM CCS International Workshop on Managing Insider Security Threats (MIST'15), Denver, Colorado, USA*. ACM, October 2015, pp. 13–20. [Online]. Available: http://dx.doi.org/10.1145/2808783.2808784

[3] F. T. Liu, K. M. Ting, and Z.-H. Zhou, "Isolation forest," in *Proc. of the 8th IEEE International Conference on on Data Mining (ICDM'08), Pisa, Italy*. IEEE, December 2008, pp. 413–422.

[4] Infosecurity Magazine, "50% job leavers steal confidential company data," http://www.infosecurity-magazine.com/view/26986/50-job-leavers-steal-confidential-company-data-/, July 2012, [Online; Accessed on December 10, 2015].

[5] S. Mathew, M. Petropoulos, H. Q. Ngo, and S. Upadhyaya, "A data-centric approach to insider attack detection in database systems," in *Proc. of the 13th International Symposium on Recent Advances in Intrusion Detection (RAID'10), Ottawa, Ontario, Canada, LNCS*, vol. 6307. Springer Berlin Heidelberg, September 2010, pp. 382–401. [Online]. Available: http://dx/doi.org/10.1007/978-3-642-15512-3_20

[6] W. Eberle, J. Graves, and L. Holder, "Insider threat detection using a graph-based approach," *Journal of Applied Security Research*, vol. 6, no. 1, pp. 32–81, 2010.

[7] M. Kandias, A. Mylonas, N. Virvilis, M. Theoharidou, and D. Gritzalis, "An insider threat prediction model," in *Proc. of the 7th International Conference on Trust, Privacy and Security in Digital Business (TrustBus'10), Bilbao, Spain, LNCS*, vol. 6264. Springer Berlin Heidelberg, August 2010, pp. 26–37. [Online]. Available: http://dx.doi.org/10.1007/978-3-642-15152-1_3

[8] F. L. Greitzer, L. J. Kangas, C. F. Noonan, and A. C. Dalton, *Identifying at-risk employees: A behavioral model for predicting potential insider threats*. Pacific Northwest National Laboratory Richland, WA, 2010.

[9] G. Magklaras and S. Furnell, "Insider threat prediction tool: Evaluating the probability of it misuse," *Computers & Security*, vol. 21, no. 1, pp. 62–73, 2001.

[10] H. Eldardiry, E. Bart, J. Liu, J. Hanley, B. Price, and O. Brdiczka, "Multi-domain information fusion for insider threat detection," in *Proc. of the IEEE 2013 Security and Privacy Workshops (SPW'13), San Francisco, California, USA*. IEEE, May 2013, pp. 45–51. [Online]. Available: http://dx.doi.org/10.1109/SPW.2013.14

[11] R. F. Mills, M. R. Grimaila, G. L. Peterson, and J. W. Butts, "A scenario-based approach to mitigating the insider threat," DTIC Document, Tech. Rep. ADA545628, 2011.

[12] A. Memory, H. G. Goldberg, and E. Ted, "Context-aware insider threat detection," in *Proc. of Workshops at the 27th AAAI Conference on Artificial Intelligence (AAAI-13), Bellevue, Washington, USA*. the Association for the Advancement of Artificial Intelligence, July 2013, pp. 44–47.

[13] W. T. Young, A. Memory, H. G. Goldberg, and E. Ted, "Detecting unknown insider threat scenarios," in *Proc. of the IEEE 2014 Security and Privacy Workshops (SPW'14), San Jose, California, USA*. IEEE, May 2014, pp. 277–288. [Online]. Available: http://dx.doi.org/10.1109/SPW.2014.42

[14] H. Eldardiry, K. Sricharan, J. Liu, J. Hanley, O. Brdiczka, B. Price, and E. Bart, "Multi-source fusion for anomaly detection: using across-domain and across-time peer-group consistency checks," *JJournal of Wireless Mobile Networks, Ubiquitous Computing, and Dependable Applications (JoWUA)*, vol. 5, no. 2, pp. 39–58, June 2014.

[15] A. Patil, J. Liu, J. Shen, O. Brdiczka, and J. Gao, "Modeling attrition in organizations from email communication," in *Proc. of the 2013 IEEE International Conference on Social Computing*

(SocialCom'13), Alexandria, Virginia, USA. IEEE, September 2013, pp. 331–338. [Online]. Available: http://dx.doi.org/10.1109/SocialCom.2013.52

[16] J. Wang, Y. Zhang, C. Posse, and A. Bhasin, "Is it time for a career switch?" in *Proc. of the 22nd International World Wide Web Conference (WWW'13), Rio de Janeiro, Brazil*. ACM, May 2013, pp. 1377–1387.

[17] J. Glasser and B. Lindauer, "Bridging the gap: A pragmatic approach to generating insider threat data," in *Proc. of the IEEE 2013 Security and Privacy Workshops (SPW'13), San Francisco, California, USA*. IEEE, May 2013, pp. 98–104.

[18] K. Sricharan, G. Gavai, D. Gunning, R. Rolleston, M. Singhal, J. Hanley, J. Liu, and O. Brdiczka, "Detecting employee churn from enterprise social and online activity data," in *Proc. of the ASE 8th International Conference on Social Computing, Stanford, California, USA*, August 2015.

[19] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. New York, NY: John Wiley and Sons, Inc., 1991.

[20] L. Breiman, "Random forests," *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.

[21] J. E. Orr, "Ethnography and organizational learning: In pursuit of learning at work," in *Organizational Learning and Technological Change*, ser. NATO ASI Series, C. Zucchermaglio, S. Bagnara, and S. U. Stucky, Eds. Springer Berlin Heidelberg, 1993, vol. 141, pp. 47–60.

Author Biography

Gaurang Gavai currently works on applications in embedded reasoning and machine learning with a focus on social network data analysis and technical infrastructure development therein. He has designed and developed systems and algorithms to both identify anomalous trends in data and also allow subject matter experts to apply their expertise by abstracting the nuts and bolts of the analyses from these tools. Prior to PARC, Gaurang was a student at the Georgia Institute of Technology where he obtained his Master's in Computer Science with a specialization in Machine Learning. He worked as a Research Assistant on developing new models of Active Learning and also interned at Citadel LLC as a Financial Technology intern along the way. Gaurang also worked as a part-time web developer during his undergraduate education in the University of Mumbai where he earned his Bachelor's in Information Technology.

Kumar Sricharan currently focuses on statistical machine learning and data mining methods for anomaly detection and pattern recognition in multivariate, temporal, and relational data. His research interests include statistics, machine learning, data mining, and signal processing with specific focus on ensemble methods and large sample estimation theory. He has particular interest in applications concerning anomaly detection and structure discovery in data. Prior to PARC, Sricharan was a research engineer at NASA Ames, where he conducted research on mining aviation data for anomalies with regard to fuel consumption efficiency and aviation safety. He also completed a R&D internship at General Motors, with research on classifying driving behavior based on statistical analysis of headway time-series data. Sricharan earned his Ph.D. in Electrical Engineering, Systems in 2012, M.A. in Statistics in 2011, and M.S. in Electrical Engineering: Systems in 2009, all from the University of Michigan, Ann Arbor. His doctoral work on efficient estimation of probability density functionals using neighborhood graphs has resulted in publications in esteemed peer-reviewed conferences and journals, and is currently nominated for the best dissertation award at the University of Michigan. He also received his B.Tech degree in Electrical Engineering from IIT Madras in 2006.

Dave Gunning directs PARC's efforts in artificial intelligence and predictive analytics focused on the enterprise. These include projects in anomaly and fraud detection, contextual intelligence, recommendation systems, and tools for smart organizations. The technical focus is on developing rich, predictive user models that capture and understand the situational context well enough to prioritize information, make recommendations, and act on the user's behalf. Dave is an experienced technology manager with an extensive background in the development and application of artificial intelligence technology. Prior to PARC, Dave was a Senior Research Manager at Vulcan Inc., a Program Manager at DARPA (twice), SVP of SET Corp., VP of Cycorp, and a Senior Scientist in the Air Force Research Labs. At DARPA, he managed the Personalized Assistant that Learns (PAL) project that produced Siri and the Command Post of the Future (CPoF) project that was adopted by the US Army in Iraq. Dave holds a M.S. in Computer Science from Stanford University, a M.S. in Cognitive Psychology from the University of Dayton, and a B.S. in Psychology from Otterbein College.

John Hanley works on large software systems deployed in situations spanning the breadth of PARC's customers, from decision support to aerospace applications. Prior to his current work in Embedded Reasoning, John contributed to a PARC and Xerox research effort aimed at retrieving documents from a personal library exactly when they become relevant. Before joining PARC, John served as Manager of Internet Services at Oracle, and as Technical Yahoo at a well-known web portal, before going on to earn his Masters in Software Engineering at Carnegie Mellon University. After collaborating with several developers of new ether probes, John brings his experience back to the home of the Digital Intel Xerox blue book, the spec that brought networking to the masses.

Mudita Singhal is a Senior Research Scientist in the Technology for Agile Organizations (TAO) Group. She has more than 12 years of experience working in interdisciplinary teams facilitating product development and heterogenous data integration, visualization & analysis. Her expertise is in the areas of visual analytics, user interfaces and applied machine learning. Her current research focuses on developing visual analytic products for Law Enforcement Agencies. She was previously the Principal Investigator of a government grant building POCs for search across large document collections. Mudita has been the product owner and made significant contributions to several publicly deployed tools, including COPASI, BRM, Pquad, CABIN and Cartoonist, as well as internally deployed POCs such as Footprints and CoRef. She has obtained her Ph.D. in Computer Science from Washington State University and a Masters in Computer Science from Virginia Tech. She is Product Management Certified from the Haas School of Business, Berkeley and has managed several government and internal grants. Additionally, she has co-authored more than 35 peer-reviewed journal and conference publications.

Rob Rolleston currently works in the area of Information Visualization. His technical interests have always been related to how people see, perceive, and interact visually with the world around them; and in turn helping others understand these mechanisms. "The world is exploding with data and information. If we are smart about how this is presented to people, there is so much we can learn about the world." Rob previously has worked in the areas of Color Management, Strategy & Planning, and assorted management roles within the Xerox Research Labs. Rob also has been an adjunct professor and instructor at Rochester Institute of Technology and is currently an instructor at Maryland Institute College of Art. He has served on the Executive Advisory Board for the New York State Center for Electronic Imaging Systems and currently serves on the Advisory Board for the Rochester Institute of Technology Center for Imaging Science. He is currently chair of the Xerox University Affairs Committee. Dr. Rolleston holds a B.S. in Computational Physics from Carnegie-Mellon University, and M.S. and Ph.D. in Optics from the University of Rochester, and a MPS in Information Visualization from Maryland Institute College of Art. He has 33 patents in the areas of color management, image processing, and virtual rendering.

A Red Team scenarios

Stealing Login Credentials
: An employee steals usernames and passwords from co-workers and emails them to an outside party.

Exfiltration Prior to Termination
: An employee is leaving the company and decides to take all of their emails and files with them.

Masquerading
: One user is masquerading as another on an unattended workstation.

Bona Fides
: Espionage volunteer prints a bona fides package and takes it to a foreign embassy.

Hiding Undue Affluence
: An employee possesses undue affluence because of ongoing espionage activity. They need to hide the existence of the money from investigators and they perform research on how to do so.

Exfil with Complex Steganography
: An employee uses steganography to hide data in an image file, then uploads that file to a website.

Insider Startup
: Three co-conspirators collude to steal company IP. They coordinate the synchronized theft of proprietary information before leaving the company.

Masquerading 2
: Subject sets up a rogue SSH server on another user's machine. They also make a copy of the local Windows password file and copy the file off over the network.

Indecent RFP
: Subject uses an inappropriate relationship with another employee to illegally influence vendor selection for a lucrative catering contract in order to obtain personal financial gain.

Credit Czech

Subject runs an illicit business trafficking in stolen credit card numbers, using the organization's IT resources. He acts as a middleman between various external purveyors of stolen numbers and a Russian operative who buys the collected numbers.

Czech Mate

Similar to Credit Czech, but the new protocol calls for twice-daily emails to Subject's Russian counterpart in order to keep the operation alive.

Naughty by Proxy

A disgruntled employee seeks revenge by logging on to her manager's computer and visiting questionable websites.