Poisson Factorization for Peer-Based Anomaly Detection

Melissa Turcotte*, Juston Moore*, Nick Heard† and Aaron McPhall*
*Advanced Research in Cyber Systems, Los Alamos National Laboratory
†Imperial College London & Heilbronn Institute for Mathematical Research, University of Bristol

Abstract—Anomaly detection systems are a promising tool to identify compromised user credentials and malicious insiders in enterprise networks. Most existing approaches for modelling user behaviour rely on either independent observations for each user or on pre-defined user peer groups. A method is proposed based on recommender system algorithms to learn overlapping user peer groups and to use this learned structure to detect anomalous activity. Results analysing the authentication and process-running activities of thousands of users show that the proposed method can detect compromised user accounts during a red team exercise.

I. INTRODUCTION

Detecting compromised user credentials and credential misuse in enterprise computer networks is an important and challenging problem. The 2016 Verizon Data Breach report states that 63% of confirmed data breaches involved stolen user credentials [1], and the 2015 Ponemon Cost of Cyber Crime Study found that malicious insiders posed the threat type which took the longest to detect and was most expensive on a per-incident basis [2]. Cyber incident responders currently rely on a combination of rule-based intrusion detection systems and human intuition to identify credential misuse. Rule-based techniques are a powerful tool, but they require specific threat signatures which can be subverted by increasingly sophisticated cyber attacks, both from external and internal actors.

Statistical anomaly detection systems offer an orthogonal defence by comparing user actions with an established behaviour model. By automatically learning and refining models of user behaviour over time, anomaly detection systems require little input from human analysts and quickly adapt to new networks upon deployment. Anomaly detection systems have gained popularity both within the academic community [3]–[5] and more recently in industry where they are commonly referred to as user behaviour analytics (UBA).

Due to the large diversity of actions that a computer user can take, predicting the probability of unseen events is a serious challenge. Incorporating peer-based analysis allows anomaly detection systems to incorporate the behaviour of a user’s peers to better learn individual profiles, and thus reduce false alarms. In industry, peer groups are often formed using human resources information, such as job title or organisation, and these global peer groups are assigned across all user activities. Learning peer groups based on observed user behaviour allows the system to form different groupings of users depending on the activity or feature of the data being analysed, and to learn each peer groups behaviour from the data.

This article proposes probabilistic recommender system algorithms for peer-based anomaly detection. Recommender systems are widely used for predicting the ‘rating’ or ‘preference’ a user will give an item, based on historical data about user activities. Collaborative filtering approaches have been shown to perform well for large collections of items, such as movie databases, and have been used for a wide range of predictive and exploratory tasks [6]. So far, collaborative filtering methods have been under-utilised for anomaly detection. Related graph-based methods have been proposed, but they do not provide a direct probabilistic interpretation [7].

The Poisson factorization model for collaborative filtering developed by Gopalan et al. [8] is utilised and two user activities are considered: the processes run by the user, and machines on which users authenticate. A recommendation system is built for each of these observed activities and allow the model to capture different peer group structures to model each activity type.

Section II reviews the model presented in [8], and Section III explains how the fitted model can be used for anomaly detection. Section IV introduces the motivating data set and presents the results of the analysis.

II. POISSON RECOMMENDATION

For $n$ users and $m$ items, let $Y \in \mathbb{R}_{+}^{n \times m}$ be a matrix of counts, where element $Y_{ui}$ is the random variable for the number of times the user $u$ invoked process $i$, or authenticated to machine $i$. These data can be modelled using a $k$-dimensional Poisson factorization model, where each item $i$ and user $u$ are represented by non-negative $k$-vector latent factors $\theta_u = (\theta_{u1}, \ldots, \theta_{uk})$ and $\beta_i = (\beta_{i1}, \ldots, \beta_{ik})$ respectively. The counts $Y_{ui}$ are assumed to follow a Poisson distribution with mean given by the dot product of the latent variables,

$$Y_{ui} \sim \text{Poisson}(\theta_u \cdot \beta_i).$$

To capture diversity in the activity levels across the user and item populations, the model incorporates hierarchical gamma priors for the latent factors [8]

$$\theta_{uj} \sim \text{Gamma}(a_u, \xi_u), j = 1, \ldots, k, \quad \xi_u \sim \text{Gamma}(a', b'),$$

$$\beta_{ui} \sim \text{Gamma}(b_i, \eta_i), j = 1, \ldots, k, \quad \eta_i \sim \text{Gamma}(c', d'),$$

so that the hyperparameters $\xi_u$ and $\eta_i$ correspond to overall activity levels for user $u$ and item $i$. 
Given an observed user-item matrix of counts $Y$, inferential interest is focused on the marginal posterior distribution $\{\theta, \beta | Y\}$, since this underpins the predictive distribution on which user-item pairs are likely to be observed in the future. Since the posterior does not have a closed-form solution, [8] uses variational inference which is an optimisation-based technique providing analytic approximations to intractable posterior distributions for complex models. The mean-field variational algorithm from [8] will be utilised here, as referenced to the code provided at 1, and the reader is referred to [8] for details.

III. ANOMALY DETECTION

The problem statement with respect to anomaly detection is to determine if the observed user-item pairs over some time period are considered normal with respect to the model learned over some training period or if they can be considered anomalous. For user $u$ and item $i$, an observed count $y_{ui}$ during a testing period is given an anomaly score equal to the upper tail probability of $y_{ui}$ given the posterior expected values of the latent factors, $\theta_u$ and $\beta_i$.

$$p_{ui} = Pr(Y_{ui} \geq y_{ui} | \theta_u, \beta_i).$$

(2)

Note that this serves as a computationally tractable approximation to the true posterior predictive upper tail $p$-value, which again does not have a closed-form solution.

Given a sequence of observed counts across items for a user, $y_{u1}, \ldots, y_{um}$, the $p$-values (2) can be combined to give an overall anomaly score for each user. Fisher’s method is commonly used to combine $p$-values obtained from independent tests into the single test statistic

$$X_u = -2 \sum_{i=1}^{m} \log(p_{ui}).$$

Under the null hypothesis that the model is correct, $X_u \sim \chi^2_m$. Outlying behaviours correspond to large values of $X_u$, and so a single combined $p$-value

$$p_u = Pr(X_u > x_u).$$

(3)

from the upper tail of $\chi^2_m$, represents a measure of surprise for each user $u$.

IV. EXPERIMENTS

A. Los Alamos National Laboratory Network Host Logs

The data set used for analysis is taken from an internal collection of host logs collected by Kent et al [9]. The data was collected over a two month period from computers running the Microsoft Windows operating system on Los Alamos National Laboratory’s (LANL) enterprise network. The data has been released and is available for download 2.

In the test month of data considered here, there are 91 known user credentials which were compromised during a month-long red team exercise within the LANL network. The aim of the analysis is to detect these compromised credentials using the Poisson factorization approach.

In particular, two features of user behaviour are analysed: the processes invoked by the user, and the computers on which users authenticate. For this analysis, interest focuses on what [8] refers to as the “implicit” data, whereby observation of the count matrix $Y$ is treated as censored [10], recording only whether $Y_{ui} = 0$ or $Y_{ui} > 0$. Two data sets are considered:

- User-Process, a binary matrix with $n = 8,786$ users and $m = 11,571$ processes. There are 360,065 observations over a 1 month training period and 385,631 observations over a 1 month test period when the red team exercise occurred. For the process names some standardisation was performed so that processes running with different versions were mapped to the same process.
- User-Authentication, a binary matrix with $n = 9,232$ users and $m = 12,750$ computers with a total of 69,697 observations over the 1 month training period and 69,526 observations over the 1 month test period when the red team exercise occurred.

Any users or items that were present in the test period and not in the training period were removed from analysis. In future work, the inference procedure should be extended to deal with the arrival of new items, such as utilising content-based recommendation algorithms as in [11].

The variational inference algorithm requires a validation set to determine convergence of the algorithm, so following [8] 1% of all training observations were set aside. Note that unlike traditional test sets used for recommender systems, observations in the test set overlap those in the training set.

B. Anomaly Detection Results

Prior parameter settings were chosen so as to maximise the posterior predictive likelihood on the heldout validation set resulting in the number of latent variables $k = 50$ for the User-Authentication data set, $k = 50$ for the User-Process data set and the prior parameters specified in (1) as $a = b = a' = c' = .5$, $b' = d' = .01$.

For evaluation of the Poisson model fit, an $N$-precision statistic is calculated for each user in the test set, with the 91 known compromised credentials removed. For each user a list of the top $N$ recommendations are generated, ordered in terms of the dot products $\theta_u \cdot \beta_i$, $i = 1, \ldots, m$. The precision for user $u$ is then the proportion of those recommendations which are subsequently observed during the test period.

Fig. 1 shows the average $N$-precision, for different $N$, across the users as a function of user activity for both the User-Authentication and User-Process data. As might be expected, the precision for users who are least active is worse than users with a higher activity level. The precision performance is better for recommending processes than authentications; a reason for this could be that the most common processes will be run by almost all users; whereas the machines which users authenticate to will be more sparse and diverse. This can be seen in Fig. 2, which compares the popularity of different processes and the machines that users authenticate on.
Fig. 2. Log-log plot of the empirical distribution of the popularity of processes and machines which users authenticate on.

Finally, Fig. 3 shows the ROC curve for the sequence of p-values (3) for the User-Process data, User-Authentication data, and the combined score, where the combined score is taken as the average of the p-values from the process and the authentication data. The number of true positives in the top N most anomalous users, as a function of N, is given in Table I. The performance is best for the User-Authentication data, indicating much better signal for this particular red-team exercise in looking at which machines users authenticated on. The goal of a red-team exercise is often to steal privileged user credentials and traverse the network, which would result in more signal in looking at machines that users authenticated on, rather than the processes they ran.

V. CONCLUSION

A collaborative filtering approach based on Poisson factorization is proposed for peer-based anomaly detection. The methodology is shown to perform well in detecting compromised user credentials from a real red team exercise. Future work is to extend the model using content-based filtering for new items observed in the network as mentioned in Section IV-A; utilising content would also enhance performance for predicting known processes as the characteristics of a process, such as its parent process or its source directory will cluster items with similar properties. One problem associated with recommender systems is that they struggle to predict items used by only one individual in the network; as can be seen in Fig. 1, this is a common attribute of these data where there are many processes and machines used by only a single user. Future research will seek to use the posterior parameters learned using the recommender system as an informative prior for a multinomial-Dirichlet model on users and items, as in [12]. The multinomial-Dirichlet distribution can provide an extra layer to model an individual user’s profile, and the informative prior from the recommender system should give better predictive probabilities for new items for the user based on the behaviour of similar users.

REFERENCES