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
† Department of Mathematics, Imperial College London
and Heilbronn Institute for Mathematical Research, University of Bristol

***Abstract*—Anomaly detection techniques for identifying compromised user credentials in an enterprise network are an important research problem, garnering much attention within industry over recent years. One important aspect of the research problem is peer-based user analysis. A method based on recommender system algorithms is proposed here, quantifying when a user activity is unlikely based on the behavior of similar users. Building several recommender system algorithms for separate user activities provides an additional advantage of allowing for different peer group structures depending on the user activity being considered.**

I. Introduction

Detecting compromised or rogue users in an enterprise network continues to be a challenging and important problem. The latest 2016 Verizon Data Breach report [1], claimed that 63% of data breaches involved stolen user credentials. Furthermore, insider threat continues to be a growing problem, and is often the hardest form of credential misuse to detect. Traditional intrusion detection techniques are largely rule-based requiring specific threat signatures, which can be ineffective when dealing with increasingly sophisticated cyber attacks, both from external and internal actors. Anomaly detection and machine learning techniques for identifying compromised users within a network have gained popularity both within the academic community [2]–[4] and more recently in industry where they are commonly referred to as user behavior analytics (UBA).

With UBA the goal is to model the behavior patterns of users within the network and then predict which activities are unlikely, both with respect to the activity of the individual user and that of their peers. Peer-based analysis is extremely important with any UBA system to both reduce false alarms and detect when users are behaving anomalously with respect to their peers. In industry, peer groups are often formed using directory groupings or human resources information and these global peer groups are assigned across all user activities. A more robust approach would be to learn peer groups based on users displaying similar behavior as observed in the data and allow for different groupings of users depending on the activity or feature of the data being analyzed and then detect anomalies with respect to the inferred peer groups.

This article proposes the use of recommender systems for peer-based anomaly detection, with separate instantiations of the algorithm for different features of the data. Recommender systems are widely used for predicting the 'rating' or 'preference' a user will give an item, based on historical data about which items the user has consumed, where *consumed* could refer to a number of actions such as rating, clicking or viewing. Collaborative filtering approaches specifically aim to predict a user preference by exploiting similarity to other users through the items that they consume. So far, the use of recommender systems for anomaly detection has been under utilized, although notably in [5] a nearest neighbor based algorithm is utilized for anomaly detection on bipartite graphs.

In this paper a specific collaborative filtering algorithm developed in [6] is employed, where a Poisson factorization model is used for recommendations. In particular, two user activities are considered: the processes run by the user, and machines on which users authenticate.

Section II reviews the model presented in [6], 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 $\boldsymbol{Y} \in \mathbb{N}_0^{n \times m}$ 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 modeled 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, [6] use hierarchical gamma priors for the latent factors

$$\begin{aligned} \theta_{uj} &\overset{\text{iid}}{\sim} \text{Gamma}(a, \xi_u), j = 1, \ldots, k, \quad \xi_u \sim \text{Gamma}(a', b'), \\ \beta_{uj} &\overset{\text{iid}}{\sim} \text{Gamma}(b, \eta_i), j = 1, \ldots, k, \quad \eta_i \sim \text{Gamma}(c', d'), \end{aligned} \tag{1}$$

so that the hyperparameters ξ_u and η_i correspond to overall activity levels for user u and item i.

Given an observed user-item matrix of counts $\boldsymbol{Y}$, inferential interest is focused on the marginal posterior distribution $[\boldsymbol{\theta}, \boldsymbol{\beta} | \boldsymbol{Y}]$, 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, [6] uses variational inference which is an optimization-based technique providing analytic approximations to intractable posterior distributions for complex models. The mean-field variational algorithm from [6] will be utilized here using the code provided at https://github.com/premgopalan/hgaprec, and the reader is referred to [6] 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, $\hat{\theta}_u$ and $\hat{\beta}_i$,

$$p_{ui} = \Pr(Y_{ui} \geq y_{ui} | \hat{\theta}_u, \hat{\beta}_i). \tag{2}$$

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, $X_u \sim \chi^2_{2m}$. Outlying behaviors correspond to large values of X_u, and so a single combined p-value

$$p_u = \Pr(X_u > x_u). \tag{3}$$

from the upper tail of χ^2_{2m} represents a measure of surprise for each user u.

IV. Analysis

A. Los Alamos National Laboratory network host logs

The data set used for analysis is taken from an internal collection of host logs over a two month period from computers running the Microsoft Windows operating system on Los Alamos National Laboratory's (LANL) enterprise network. The data are available from http://csr.lanl.gov/data/cyber1/ and [7] provides a detailed description of the data.

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 behavior will be analyzed: the processes invoked by the user, and the computers in the network on which they authenticated. For this analysis, interest focuses on what [6] refers to as the "implicit" data, whereby observation of the count matrix $\boldsymbol{Y}$ is treated as censored [8], 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 standardization was performed whereby version numbers were removed from the process name 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 utilizing content-based recommendation algorithms as in [9].

The variational inference algorithm requires a validation set to determine convergence of the algorithm, so following [6] 1% of all training observations were set aside. Note that unlike traditional test sets used for recommender systems, observations in the test set will overlap those in the training set.

B. Anomaly detection results

Prior parameter settings were chosen so as to maximize the posterior predictive likelihood on the held out validation set resulting in the number of latent variables $k = 10$ 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.

Figure 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 much 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 much more sparse and diverse. This can be seen in Figure 2, which compares the popularity of different processes and the machines that users authenticate on.

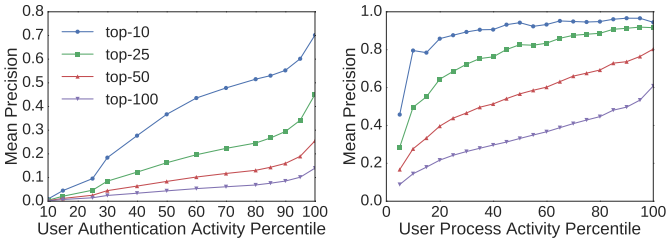

Fig. 1. Mean precision for users with varying levels of activity

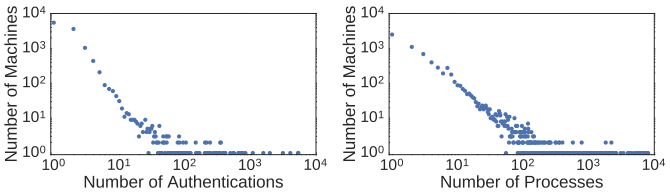

Fig. 2. Log-log plot of the empirical distribution of the popularity of processes and machines which users authenticate on.

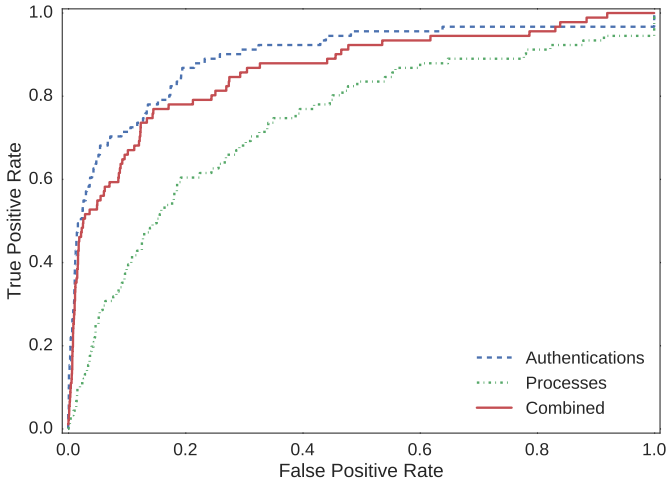

Fig. 3. Red team ROC curve for the processes, authentications and combined scored.

Finally, Figure 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 of from the process data 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.

C. Conclusion

A collaborative filtering approach based on Poisson factorization is proposed for peer-based anomaly detection of users in an enterprise network. The methodology is shown to perform well in detecting compromised user credentials. Future work is to extend the model using content-based filtering for new items observed in the network as mentioned in Section IV-A; utilizing 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 Figure 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 [10]. 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 behavior of similar users.

TABLE I

	Top N detections				
True positives	5	10	25	50	100
Processes	0	1	1	3	4
Authentications	3	4	12	17	25
Combined	1	2	5	9	22

References

[1] Verizon, "2016 data breach investigations report," Tech. Rep., 2016.

[2] S. H. Oh and W. S. Lee, "An anomaly intrusion detection method by clustering normal user behavior," *Computers and Security*, vol. 22, no. 7, pp. 596 – 612, 2003.

[3] B. Viswanath, M. A. Bashir, M. Crovella, S. Guha, K. P. Gummadi, B. Krishnamurthy, and A. Mislove, "Towards detecting anomalous user behavior in online social networks," in *23rd USENIX Security Symposium (USENIX Security 14)*. USENIX Association, 2014, pp. 223–238.

[4] S. Mukkamala, G. Janoski, and A. Sung, "Intrusion detection using neural networks and support vector machines," in *Neural Networks, 2002. IJCNN '02. Proceedings of the 2002 International Joint Conference on*, vol. 2, 2002, pp. 1702–1707.

[5] J. Sun, H. Qu, D. Chakrabarti, and C. Faloutsos, "Neighborhood formation and anomaly detection in bipartite graphs," in *Fifth IEEE International Conference on Data Mining (ICDM'05)*, 2005, p. 8.

[6] P. Gopalan, J. M. Hofman, and D. M. Blei, "Scalable recommendation with poisson factorization," *CoRR*, vol. abs/1311.1704, 2013.

[7] A. D. Kent, "Cyber-security data sources for dynamic network research," in *Dynamic Networks in Cybersecurity*. Imperial College Press, 2015.

[8] W. H. Greene, "Censored data and truncated distributions," *Available at SSRN 825845*, 2005.

[9] P. K. Gopalan, L. Charlin, and D. Blei, "Content-based recommendations with poisson factorization," in *Advances in Neural Information Processing Systems 27*. Curran Associates, Inc., 2014, pp. 3176–3184.

[10] M. J. M. Turcotte, N. A. Heard, and A. D. Kent, "Modelling user behaviour in a network using computer event logs," in *Dynamic Networks in Cybersecurity*. Imperial College Press, 2016.