Modeling and Detecting Anomalous Topic Access

Siddharth Gupta[1], Casey Hanson[2], Carl A Gunter[3],
Mario Frank[4], David Liebovitz[5], Bradley Malin[6]
[1,2,3,4]Department of Computer Science, [3,5]Department of Medicine, [6]Department of Biomedical Informatics
[1,2,3]University of Illinois at Urbana-Champaign, [4]University of California, Berkeley,
[5]Northwestern University, [6]Vanderbilt University

***Abstract*—There has been considerable success in developing strategies to detect insider threats in information systems based on what one might call the *random object access model* or ROA. This approach models illegitimate users as ones who randomly access records. The goal is to use statistics, machine learning, knowledge of workflows and other techniques to support an anomaly detection framework that finds such users. In this paper we introduce and study a *random topic access model* or RTA aimed at users whose access may be illegitimate but is not fully random because it is focused on common semantic themes. We argue that this model is appropriate for a meaningful range of attacks and develop a system based on topic summarization that is able to formalize the model and provide anomalous user detection effectively for it. To this end, we use healthcare as an example and propose a framework for evaluating the ability to recognize various types of random users called *random topic access detection* or RTAD. Specifically, we utilize a combination of Latent Dirichlet Allocation (LDA), for feature extraction, a k-nearest neighbor (k-NN) algorithm for outlier detection and evaluate the ability to identify different adversarial types. We validate the technique in the context of hospital audit logs where we show varying degrees of success based on user roles and the anticipated characteristics of attackers. In particular, it was found that RTAD exhibits strong performance for roles are described by a few topics, but weaker performance when users are more topic-agnostic.**

***Keywords*—: Data Mining, Anomaly Detection, Healthcare Security, Electronic Health Records, Access Logs, Insider threats**

I. Introduction

Consider the following problem sometimes encountered in hospital emergency departments: an initial user logs into a terminal and other users access patient records with the login of this user. Such situations occur in many enterprise contexts and they are anathema to security specialists, but they arise when users do not consider the security risks to be great enough to merit the inconvenience imposed by best practice security measures (like unique accounts that are never shared). How could an audit log review reveal an abuse like this? Much of the current work on anomaly detection in collaborative information systems has focused on a *Random Object Access (ROA)* where the object in question is a patient's electronic medical record. In this model, illegitimate users are modeled as ones who access files randomly (e.g., [1]). This is a plausible model since illegitimate accesses may look random because they are often based on features that lie outside the business activities of the organization managing the information system (e.g., accessing the record of a famous actor with an ordinary disease). However, the model may not apply to the open terminal case since the users do not make random accesses; they make accesses that are appropriate to their roles while ascribing these to a user for whom the action is inappropriate. Consider also a related problem, where a user does tasks that should be done by another user. This behavior may be inappropriate, but it probably will not look like the access to a random set of patient records.

In this paper, we overcome this conceptual limitation and propose anomaly detection strategy based on the *Random Topic Access (RTA)* model. RTA models illegitimate user behavior as random accesses to "topics" rather than objects. In this case, a topic is an idea derived from the field of machine learning, where, for instance, an algorithm is used on a corpus of articles to derive groups of words that often occur together and represent key topics of the articles. For this paper, we apply these techniques to a hospital information system, which is a canonical collaborative information system that has received increasing attention in recent years. Specifically, in our system, patients play a role similar to articles while data in the patients' medical records play a role similar to the words in the articles. The main premise of this approach is to derive a model of the topics in the organization (e.g., common groups of properties of patients) and use these to characterize the interests of users, who can be viewed as the readers of these topics. Then, an RTA user is one who accesses a collection of topics at random. We note that this is subtly, though critically, different from an ROA user, whose access to *patients* is random. The RTA model is useful for detecting anomalous users who are systematic in using topic about patients in the hospital, but are potentially unusual in the combination of topics they access. For instance, consider a group of nurses who work in the stroke unit and commonly access patient records with neurological diagnoses, but among whom there is one nurse that also accesses records with obstetrics-related activities. This may not be indicative of illicit activity, but it might be appropriately flagged as anomalous.

Our specific technique, which we call *Random Topic Access Detection (RTAD)*, uses Latent Dirichlet Allocation (LDA) [2] to define topics and a k-nearest neighbor (k-NN) matching algorithm to detect users who are anomalous in the RTA context. The principle novelty of the work is not in the use of these specific techniques for anomaly detection (since similar techniques have been used in other contexts), but rather it is the idea that detection should target RTA users rather than ROA users. A particular insight is that RTA users can be characterized along a spectrum based on the likely features of their behavior, ranging from the tendency to select few topics to many topics. This approach has the advantage of being more general in comparison to simulating open terminals, masquerading, or other attack modes directly.

We evaluate our methodology in the context of hospital

audit logs where we ran LDA on four months of patient records and created a list of hospital topics that enable each patient to be described as class of topics (e.g., topics based on diagnoses, medications, procedures, locations, or services). From the results of LDA and user-patient access information in the audit logs, we derived topics that characterize users who access these patients and the hospital-assigned roles of these users. We focused our attention on five roles and five different kinds of users, allocated between the two extreme cases where users favor a few topics i) strongly and ii) weakly. For each case we compute the Area Under Curve (AUC) from Receiver Operating Curves (ROCs) for detecting RTA users. The results indicate that the effectiveness of RTAD varies with the class of topics. Given the AUC values for all class of topics for each of the five roles, we find that RTAD works better for each role than the collection of all users when they are described by a few topics. However, the performance declines for users that are more topic-agnostic.

The paper is organized as follows. In Section II we present related work on insider threat detection for hospitals and how techniques have been validated using the ROA. In Section III we describe the data set and how to derive topics in Section IV. We formalize the RTA model in Section V. We then apply this analysis to anomaly detection for five major hospital roles and the collection of all users who accessed patient records during a several month timespan in Section VI. Finally, we provide discussion and conclusions in Section VII.

II. Related Work

There is a significant amount of research into anomaly detection more generally. For instance, Das et al. [3] uses association rule mining and Bayesian approaches to discover outliers in categorical datasets. This work is limited in that it does not account for heterogeneous datasets such as electronic medical record (EMR) systems and the associated user-level access logs. There is also a good body of work (e.g., [4], [5], [6]) on threat detection models for masquerading users. Masquerading corresponds to people who actively try to portray themselves as legitimate users. They often mimic the behavior of trusted users and then deviate to perform unexpected activities. These techniques provide an interesting contrast with the ones we present below. Garg et al. [6] tried to create threat detection models for masquerading users in GUI-based systems by extracting relevant features through singular value decomposition (SVD) and used supervised learning to classify anomalous behavior. On the other hand, Wang et al. [4] and Maxion [5] extend research in masquerade detection using UNIX commands issued by the users and applied supervised learning mechanism to predict the anomalous users. Our approach differs from theirs in two primary ways. First, while [6] only includes the activity of three different users, we have annotated information for thousands of users and patients collected over a four year span. Second, these approaches generally do not scale well for large access logs in a dynamic environment such as healthcare. Also, we believe that the latent topic space provides more semantically coherent and intuitive features than those obtained through SVD.

When studying collaborative information systems, it is ideal to evaluate the effectiveness of an intrusion detection system with actual intruders. In this regard, [7] introduced a supervised learning approach, where the accesses to an electronic medical record (EMR) system deemed suspicious by administrative privacy officials were used to train a classifier that achieved reasonable results. This method was further refined in [8] through a semi-automated process of signature filtering. These methods were shown to exhibit excellent performance in the detection of a holdout set of suspicious accesses. Yet, despite the merits of these studies, there are several major concerns. First, anomalies are by definition not a class, so it is not clear if the expert analysts in such reports capture all of the potential problems. Second, the aforementioned approaches do not scale easily because every setting is different and human experts are not always available.

As a result, most auditing research with respect to EMRs is validated by detecting anomalous users and anomalous access patterns specifically for ROA type models, where the evaluation is conducted by synthetic users accessing random patients. Chen et al. [9] introduced CADS (Community Anomaly Detection System), which leverages the team-based nature of users in collaborative information systems. CADS extracts patterns of users in the form social networks, which are basically derived from singular value decompositions (SVDs) of user-patient access networks. CADS then performs anomaly detection by determining how far a user is from its k nearest neighbors in the resulting eigenspace. This approach was subsequently extended into MetaCADS [1], which accounts for the similarity of patients based on meta-information, such as their diagnoses. While CADS was designed to find anomalous users, SNAD (Specialized Network Anomaly Detection), was designed to find specific suspicious accesses of the users. This approach focuses on the local access network for a specific patient and assesses if the network is significantly different when removing a particular user from the group. For these methods, the evaluation centers on the recovery of users generated via the ROA model. The approaches were shown to have high AUC values when ROA users were mixed into the access logs of a large academic medical center. By contrast, our analysis with RTAD considers subpopulations of users by role based on RTA. We wish to point out that while these methods utilize dimensionality reduction, the resulting features only group dimensions together according to their contribution to the variance to the system, not their semantic coherence.

Beyond anomaly detection, one may view audit logs as providing a window into better access control, a view expressed in Experience-Based Access Management (EBAM) [10]. For instance, one can focus on positive explanations of accesses rather than anomalies, as explored in Explanation-Based Auditing System (EBAS) [11]. This system is based on the assumption that employees are responsible for a predictable collection of diagnoses based on their departments so their access to records of patients with these diagnoses are explained and hence not anomalous. The effectiveness of this technique is validated by an ROA model where users who randomly access patient records are added to the hospital and used to measure accuracy.

III. Data For This Study

The dataset for this study was derived from the Cerner Powerchart EMR system in place at Northwestern Memorial Hospital (NMH). It consists of all user accesses, or audit logs,

made over a four month period, in addition to EMR data for patients admitted in this timeframe. All data was de-identified for this study in accordance with the Safe Harbor standard of the HIPAA Privacy Rule.[1] As a noise reduction measure, we filtered all outpatient entries, focusing only on patients who stayed a significant amount of time at the hospital (i.e, more than 24 hours). We further removed all patients younger than 17 years old (about 9.1% of the records) from the dataset, as patients from this age group tended to have sparsely populated records. The final dataset consisted of 4.9 million accesses made by 7932 users to 14606 patients.

EMR data was accumulated with respect to given hospital visits for patients at the hospital, referred to as an encounter; however, this data does not attribute specific information in the record of a patient to the authoritative user, only permitting the association of a group of users to a patient. To prevent associating certain patient features to non-relevant users, we consider every new encounter to be a new patient. The following subsections aim to elucidate the key differences between these two distinct datasets.

A. Audit logs

Audit logs consist of user accesses to specific patients, logging one of 30 possible services and one of 49 possible hospital locations for the patients. Table I provides basic statistics summarizing this dataset. In our analysis, we typically combine service and location into a single combined dimension called service/location, due to the relative small size of these dimensions compared to those in the EMR data set.

B. EMR

EMR data consists of patient-encounter records, with each record corresponding to various diagnoses, procedures, and medications. A given dimension (e.g., diagnoses) is a binary vector, with each bit in the vector representing the presence or absence of a particular feature (i.e., a specific diagnosis code for diagnoses). A feature in this case is a specific value in a dimension. Diagnosis features, for instance, are characterized by the lowest level of the ICD-9 code hierarchy, with 4543 unique codes [12]. Procedure features are similarly defined via ICD-9-CM codes, albeit with less code words (1237). Medication features are defined with respect to RxNorm, a normalized naming system for generic and branded drugs [13]; in total, there are 642 codes. Table II provides basic statistics on the various dimensions. In addition, it is important to note the lack of patient-user information in the EMR data. Utilizing the audit logs, it is possible to associate particular users with certain patient-encounters. However, since many users may access a given patient-encounter, it is impossible to know exactly what diagnosis, medication, and/or procedure a specific user contributed without further analysis of the clinical narrative in the medical record.

TABLE I. SUMMARY STATISTICS FOR THE AUDIT LOGS

Attribute	Value
Duration of Audit Logs	4 months
Distinct Accesses	4979465
Distinct Patients	12488
Distinct Patient-Encounters	14606
Distinct Users	7932
Average Patients Accessed per User	115
Average Accesses of Patient	340

TABLE II. SUMMARY STATISTICS FOR THE PATIENT RECORDS

Attribute	Value
Distinct User Roles	156
Distinct Patient locations	49
Distinct Patient services	30
Distinct Patient diagnoses	4543
Distinct Patient procedures	1237
Distinct patient medications	996

IV. TOPIC MODELING

RTA entails modeling users as probability distributions over topics. These topics are defined with respect to the dimensions introduced in Section III. User characterization in terms of the dimensions in EMR and audit log data is independent of the characterization of other users in the system. While this can be informative, it is of limited power, as knowledge of the user does not convey any knowledge about how that user behaves in the context of the system. We utilize a latent topic framework because modeling users as distributions over topics enables several modeling advantages over alternative conceptualizations. First, users can be summarized in a semantically coherent way with respect to the entire population. Topic modeling can provide a concise description of how a user behaves in the context of his peers and the meaning of that behavior. Second, the latent topic framework provides a mechanism for user simulation. By modeling users as samples from a Dirichlet distribution over topic multinomials, the space of user behaviors is significantly larger than alternative approaches. Third, as will be elaborated upon on Section V, this framework enables more explicit control over the type of adversary.

While many algorithms for extracting and modeling topics exist, we adopt latent Dirichlet allocation (LDA) for the current version of RTAD. LDA is a generative model that characterizes documents in a corpus as multinomials over a set of latent topics [2]. These latent topics are modeled as multinomials over the words in a corpus. In this manner, topics act as summaries of the different themes pervasive in the corpus, while documents are characterized with respect to these summaries. A d dimensional multinomial is sampled from a d dimensional Dirichlet distribution to derive a particular topic distribution. This can be thought of as sampling from a $d-1$ simplex (probability space) controlled by α, the concentration parameter of the Dirichlet. Specifically, α controls where on the simplex the multinomials are likely to be sampled. As $\alpha \to 0^+$, the probability density is pushed towards the edges of the simplex, favoring multinomials heavily biased towards few topics. As $\alpha \to 1$, the probability density becomes more

[1]This included pseudonymizing all patient and user ID's with random values, random date shifting in a -365 day window, removing all geocodes for patient's home residence and recoding all patients over 89-years old as 89+.

evenly distributed around the simplex, making any multinomial equally probable of being selected. Finally, as $\alpha \to +\infty$, the probability density is pushed towards the center of the simplex, favoring multinomials equally biased towards all topics. The number of topics, k, and the concentration parameter, α, are defined a priori. In the experimental setting below, we use α to generate users at five different levels of topic-concentration, ranging from users whose actions are described by each topic contributing almost equally to users described by a single topic only.

To model users, we generate topics for diagnosis, medication, procedure, and service/location. Additionally, we generate topics over the superset of all these dimensions, referred to as the mixed-bag (i.e., LDA was run on the mixture of features taken together from each dimension to form a single dimension), and the concatenation of the topic vectors of these different dimensions, referred to as a combined-bag (i.e., the resulting topics from each dimension were concatenated to form a single dimension). The logic behind modeling users in terms of the mixed-bag is to determine if the combinations of dimensions are more informative than independent dimensions. Likewise, we considered the combined-bag to see if a naïve concatenation of these different information types is comparable to or favorable to the mixed-bag. Table III provides a visualization for topics across these different dimensions. Coherent topics are chosen for the service/location, mixed-bag, diagnosis, and medication dimensions; for each dimension the top 3 most probable features are displayed. We observed that there is a strong bias in these distributions towards women's health, with a specific focus on birth, demonstrating the efficacy and power of LDA to capture relevant semantic summarizations.

To set the number of topics, we utilized the perplexity measure, which is designed to assess the effectiveness of different topic numbers. The perplexity measure is an estimation of the expected number of equally likely features in the population, such that by minimizing perplexity we maximize the topic variance captured by the system [2]. We performed perplexity analysis for each topic distribution and the number of topics corresponding to the minimum perplexity is shown in Table IV. In future, we plan to investigate the general issue of expert opinion on the topics. It should be noted that, with respect to EMR derived dimensions, LDA was performed on patients. We use the following section to investigate how the patient topic probabilities (derived from LDA) are translated in terms of users.

A. User Typing

In the audit logs, users are summarized by performing LDA on the aggregation of their activities (i.e., each user is represented as a set of features he has accessed in the audit log). Similarly for the EMR data, the dimensions are provided with respect to patients, not users. Patients can be associated with certain users through cross-referencing users from the audit logs to the patients they access in the EMR dataset. After running LDA on patient dimensions, a user can be specified with respect to a dimension d via the following equation.

$$T(u,d) = \frac{1}{|P(u)|} \cdot \sum_{p \in P(u)} T(p,d) \tag{1}$$

Note that, $T(u,d)$ denotes the topic distribution of user u along dimension d, $P(u)$ denotes the set of patients u accesses, and $T(p,d)$ refers to the particular topic distribution of a patient.

TABLE III. THE TOP 3 FEATURES FOR THE CANDIDATE TOPICS

medication	**Prob**	**diagnosis**	**Prob**
Ibuprofen	0.158	Single Liveborn	0.193
Oxytocin	0.158	Cesarean Delivery	0.124
Docusate	0.148	Antepartum Condition	0.06
service	**Prob**	**mixed-bag**	**Prob**
Obstetrics	0.288	Single Liveborn	0.061
Labor & Delivery	0.286	Labor & Delivery	0.060
Prentice 11	0.283	Ibuprofen	0.060

V. RANDOM TOPIC ACCESS MODEL

The RTA model is a framework for describing anomalous users in terms of random topics, as opposed to random access patterns. Randomness in this sense can take on many subtle definitions. Within this framework, we argue that certain types of attackers can be elegantly and accurately synthetically generated. We will proceed with a discussion of these types of anomalous users followed by a review of our implementation of the RTAD framework. Although we frame our arguments with respect to the Dirichlet distribution, we do not necessarily assume it is the only mechanism for generating multinomials. However, considering the Dirichlet distribution is the conjugate prior of the multinomial, we believe it appropriate to convey our argument with respect to the concentration parameter, α.

1) Directed or Masquerading User: $\alpha < 1$: The first type of user the RTA model is capable of capturing is the directed, or masquerading, user. In this scenario, an anomalous user of some specialty gains sole access to the terminal of another user in the hospital. In this sense, the anomalous user is masquerading as the real user, making accesses related to his specialty while logged in as another user. The topics ascribed to the anomalous user's access patterns should differentiate this user from the real user. While these topics may be ordinary with respect to the hospital population, they could be deviant with respect to the population of users who are similar to the real user. The anomalous user in this case could be sampled from a Dirichlet with $\alpha < 1$ because real users are assumed to be strongly biased towards a set of few topics. Given a real user with an typical topic distribution, it is highly probable

TABLE IV. TOPIC SUMMARY

Dimension	**# Of Topics**
diagnosis	25
procedure	25
medication	25
service	20
mixed-bag	40
combined-bag	95

that simulating random users will result in anomalous users not biased towards the same topic as the real user.

2) Purely Random User: $\alpha = 1$: The second type of user the RTA model can handle is the purely random user. This type of user is characterized by completely random behavior, with little semantic congruence to the hospital setting. This is the ideal form of randomization that ROA models aspire to capture. However, because ROA models preferentially sample randomly from the data, it would be expected that not all random behaviors would be realized. By generating random users from a Dirichlet with $\alpha = 1$, any type of random user can be generated. This has the useful property of allowing the system to be tested against input that does not exist in the data.

3) Indirect User: $\alpha > 1$: The third type of user modeled by RTA is the indirect user. This user type resembles an even blend of the topics of many specialized users. The best analogy in the hospital setting is the open terminal problem. In this scenario, a user leaves the access to his terminal open and users of different specializations log in and make accesses under this users account. As a result, the logged-in user resembles a sort of average of these different users who have high probability for certain topics. The anomalous user can best be modeled with $\alpha > 1$ in the Dirichlet distribution. This would result in sampling preferentially from the middle of the simplex, where topic probabilities are seemingly unbiased to every topic.

VI. Analysis

Our RTA framework, RTAD, consists of 1) running LDA on the entire population of users, 2) typing users with respect to their accesses and patients, 3) identifying the top 5 most populated user roles, and 4) injecting anomalous users into each role at a 5% mix rate for various α settings: 0.01, 0.1, 1, 10, 100. Utilizing a simple k-NN algorithm, for each of the 5 most populated users and each α, we generated AUCs from the corresponding ROC curves generated by a simple linear classifier, utilizing the distance ratio in k-NN for each point as a moving threshold. The k in k-NN varied from 2 to 20 and, for each role, all feature topics were evaluated. For the purposes of comparison, we performed the same analysis for each of the individual α values on the population of users across all roles. We then compared the best AUC for each role and α.

Figures 1 and 2 show the best AUCs for each role-α combination as well as the best AUC across all α in the entire population. For masquerading users (i.e., $\alpha < 1$), the resulting AUCs are very large, especially for highly specialized users ($\alpha = 0.01$). This is expected because, when the synthetic users are driven to the edge of the simplex, it is highly probable they will not be biased towards the same topic as the majority of the users in a role. As a consequence, they will approach the maximum distance that can be achieved on the simplex and will appear more varied with respect to the users in the role. As the system transitions to more purely random users, the resulting AUCs suffer somewhat for all roles, except for NMH Physician Office - Computerized Physician Order Entry (CPOE). Analysis of this role showed the system inverting itself such that anomalous users appear more clustered on the topic simplex than actual users. This type of inversion was evident when evaluating the system against indirected users

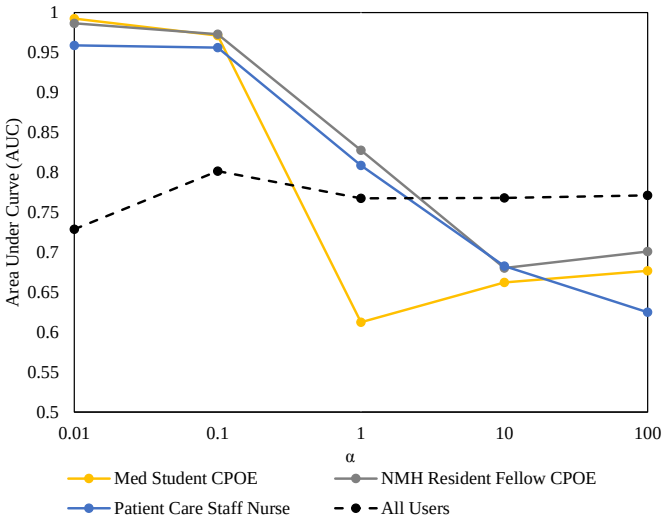

Fig. 1. The best AUC across all evaluated dimensions is plotted for each role performing badly for $\alpha > 1$.

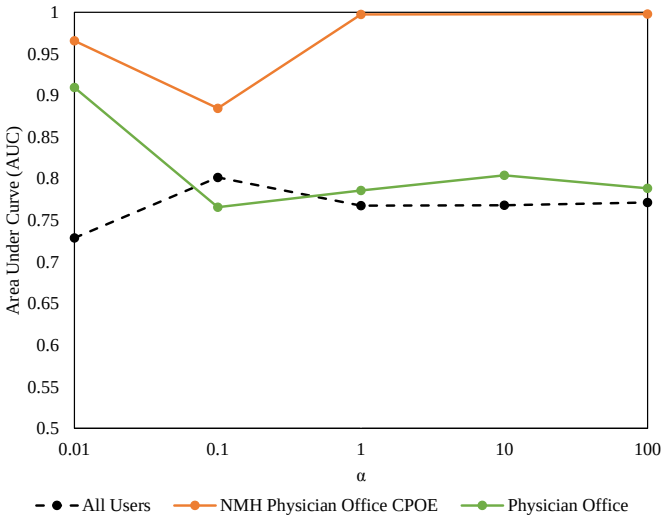

Fig. 2. The best AUC across all evaluated dimensions is plotted for each role performing well or near average for $\alpha > 1$.

as well, which matches expectation as indirected users (for increasing levels of α) show little variation between each other. Thus, it is expected that for increasing levels of α, indirected users will appear more clustered than actual users. With respect to the baseline, utilizing semantic role information the system for directed users and generally performs as well or better for purely random users. The detection performance suffers for some roles tested against undirected users compared to the baseline, but this discrepancy is intuitive in the context of k-NN. This is because the simplex is more populated in the baseline case. As such, that there is a higher likelihood of local clusters of users across different roles.

Our findings regarding the response of the Medical Student CPOE role for the RTA framework also make intuitive sense. Medical students typically undergo rotations where they specialize in a particular area of medicine for a fixed amount of time. As a result, over the 4 month sampling interval, have integrated many (and possibly very different) kinds of

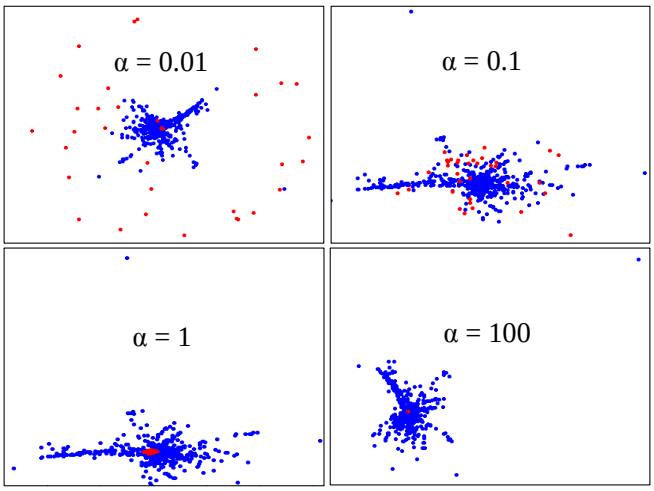

Fig. 3. Projection of NMH Resident Fellow CPOE users (blue) and random users (red) from the topic simplex to the euclidean plane, via Multidimensional Scaling (MDS). This procedure attempts to preserve relative distances between pairs of users in the high dimensional topic space in the euclidean plane. As such, the axes are unimportant - only the relative distance between points is significant.

accesses into their history. As a result, it is expected that this role has partially random behavior and is thus the most similar to the purely random user, which will yield a lower detection performance. This is intuition is further supported by the observation that as anomalous users become more tightly clustered, the detection rate in the context of the Medical Student CPOE role improves.

Our results indicated that there is no clear advantage of RTAD models based on mixed information versus concatenated vectors or mixed-bag topics versus combined-bag topics. The best topic dimensions selected for each role-α combination varied considerably and no discernible trend was detected from this small dataset.

A visualization of the effect of the α value on one of the roles is shown in Figure 3. Utilizing classical multidimensional scaling for dimensionality reduction, we graph the projections of the high dimensional topic space for NMH Physician Fellow CPOE such that the pairwise distances between users is preserved. Real users are shown in blue while anomalous users are shown in red. By comparing the top plots to the bottom plots, it can be seen that $\alpha < 1$ results in a greater amount of dispersion with respect to real users than $\alpha \geq 1$. As α increases, the random users become more clustered, making the anomaly detection task more difficult.

VII. Discussion

This paper demonstrated there is a lack of coverage in the existing methodology for evaluating security models utilizing random users. The classical technique, referred to as the Random Object Access (ROA) model, characterizes atypical behavior as random access patterns to various objects in the dataset; this approach has the distinct disadvantage of constraining the model to the particular dataset under analysis, preventing the model from imagining and evaluating a richer space of conceivable attackers. Utilizing latent topics models such as *Latent Dirichlet Allocation (LDA)*, the *Random Topic Access (RTA)* model provides greater coverage of the different types of attackers by generating synthetic users directly from a topic simplex, as opposed to data. In this manner, the dataset can be thought of as being a sample from a larger, unseen population distribution. Transformation to the topic domain may not allow the realization of new types of real users, but it enables the system to be evaluated against potentially unseen adversaries. Additionally, we posited some plausible adversarial archetypes with respect to the α parameter, which controls the distribution on the simplex. Future work along these lines includes carefully controlled experimental validation of these different types of adversaries in hospital settings, as well as investigating the efficacy of integrating labeled role information for users into the LDA component of the RTAD framework. Also, we plan to investigate our assumption that the latent topic space provides more semantically coherent and intuitive features than those obtained through SVD.

Acknowledgements

This research was supported by grants CCF-0424422 and CNS-0964063 from the NSF, R01-LM010207 from the NIH, and HHS-90TR0003/01 from the ONC but the views in the paper are those of the authors only. The authors thank You Chen of Vanderbilt University for insightful discussions during this research.

References

[1] Y. Chen and B. Malin, "Detecting anomalous insiders in collaborative information systems," *IEEE Transactions on Dependable and Secure Computing*, vol. 9, no. 3, pp. 332–344, May 2012.

[2] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *J. Mach. Learn. Res.*, vol. 3, pp. 993–1022, 2003.

[3] K. Das, J. Schneider, and N. D., "Anomaly pattern detection in categorical datasets," *KDD*, pp. 169–176, 2008.

[4] K. Wang and S. J. Stolfo, "One-class training for masquerade detection," *3rd IEEE Conference Data Mining Workshop on Data Mining for Computer Security*, 2003.

[5] R. A. Maxion, "Masquerade detection using enriched command lines," *International Conference on Dependable Systems and Networks*, pp. 5–14, 2003.

[6] A. Garg and R. Rahalkar, "Profling users in gui based systems for masquerade detection," *IEEE Workshop on Information Assurance United States Military Academy*, pp. 48–54, 2006.

[7] A. Boxwala, J. Kim, J. Grillo, and L. Ohno-Machado, "Using statistical and machine learning to help institutions detect suspicious access to electronic health records," *Journal of the American Medical Informatics Association*, vol. 18, no. 4, pp. 498–505, 2011.

[8] J. Kim, J. M. Grillo, A. A. Boxwala, X. Jiang, R. B. Mandelbaum, B. A. Patel, D. Mikels, S. A. Vinterbo, and L. Ohno-Machado, "Anomaly and signature filtering improve classifier performance for detection of suspicious access to ehrs," *Journal of the American Medical Informatics Association*, p. 723731, 2011.

[9] Y. Chen and B. Malin, "Detection of anomalous insiders in collaborative environments via relational analysis of access logs," pp. 63–74, 2011.

[10] C. A. Gunter, D. M. Liebovitz, and B. Malin, "Experience-based access management: A life-cycle framework for identity and access management systems," *IEEE Security & Privacy Magazine*, vol. 9, no. 5, September/October 2011.

[11] D. Fabbri and K. LeFevre, "Explaining accesses to electronic medical records using diagnosis information," *Journal of the American Medical Informatics Association*, vol. 20, no. 1, pp. 52–60, 2013.

[12] [Online]. Available: http://www.cdc.gov/nchs/icd.htm

[13] [Online]. Available: https://www.nlm.nih.gov/research/umls/rxnorm/