MODELING AND DETECTING ANOMALOUS TOPIC ACCESS IN EMR AUDIT LOGS

BY

SIDDHARTH GUPTA

THESIS

Submitted in partial fulfillment of the requirements
for the degree of Master of Science in Computer Science
in the Graduate College of the
University of Illinois at Urbana-Champaign, 2013

Urbana, Illinois

Adviser:

Professor Carl A. Gunter

Abstract

Recent use of Electronic Medical Records in the hospitals has raised many privacy concerns regarding confidential patient information which can be accessed by various users in the hospital's complex and dynamic environment. There has been considerable success in developing strategies to detect insider threats in healthcare information systems based on what one might call the *random object access model* or ROA. This approach models illegitimate users who randomly access records. The goal is to use statistics, machine learning, knowledge of hospital workflows and other techniques to support an anomaly detection framework that finds such users.

In this work we introduce and study a *random topic access model*, RTA, aimed at the users whose access may well be illegitimate but is not fully random because it is focused on common hospital 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 for it. We also propose a framework for evaluating the ability to recognize various types of random users called *random topic access detection*, or RTAD. The proposed RTAD framework is an unsupervised detection model which is a combination of Latent Dirichlet Allocation (LDA), for feature extraction, and a k-nearest neighbor (k-NN) algorithm for outlier detection. The analysis is done on the dataset from Northwestern Memorial Hospital which consists of over 5 million accesses made by 8000 users to 14,000 patients in a four month time period. Our results show varying degrees of success based on user roles and the anticipated characteristics of attackers and evaluate the ability to identify different adversarial types relevant to the hospital ecosystem.

To my parents, for their love and support.

Acknowledgments

I would like to thank many people who have helped me through the completion of this thesis. First and foremost, my advisor Professor Carl A. Gunter whose patience and kindness, as well as his academic experience, have been invaluable to me. Only under his guidance, I was introduced to the current challenges in HIT which further motivated me to work in this field. I am also extremely grateful to Professor Bradley Malin from Vanderbilt University, David M. Liebovitz from NMH, Casey Hanson from UIUC and Mario Frank from UC Berkeley for being wonderful collaborators. I cherish the opportunity to learn from them and gain important insights during our weekly discussions and project meetings.

I wish to thank my parents for their unconditional love, concern, support and strength throughout my life. It would have been impossible for me to have successfully made so far without their blessings. In addition, I have also been lucky to have the love of my relatives and cousins all this time. Special thanks to my aunt Achla Marathe for always being my closest guide and a well-wisher. Also, the love of my grandparents has inspired me to push my boundaries and make them proud.

I have enjoyed working with my teammates and also made great friends along the way. I am lucky to have Casey as a friend, colleague and co-author, Ting Wu for providing me strength and company during my times of stress. Fardin Abdi, Mariyam Khalid, Nikita Spirin and Siavash Bolourani for inspiring and guiding me to the right path in the time of ambiguities. Last but not the least, special thanks to my best friends Rajat Singhal, Shivin Kinra, Deepali Singh and Somya Aggarwal for being my companions in this journey.

This research was supported by grant HHS-90TR0003/01 from the ONC.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

CDC	Centers for Disease Control and Prevention
EHR	Electronic Health Record
EMR	Electronic Medical Record
HHS	Health and Human Services
HIPAA	Health Insurance Portability and Accountability Act
HIT	Health Information Technology
ICD-9	International Classification of Diseases, Ninth Revision
ICD-9-CM	International Classification of Diseases, Ninth Revision, Clinical Modification
KNN	K-Nearest Neighbor
LDA	Latent Dirichlet Allocation
MDS	Multi Dimensional Scaling
NIH	National Institutes of Health
NLM	National Library of Medicine
NMH	Northwestern Memorial Hospital
NSF	National Science Foundation
ONC	Office of the National Coordinator
ROA	Random Object Access
RTAD	Random Topic Access Model
UMLS	Unified Medical Language System

Chapter 1

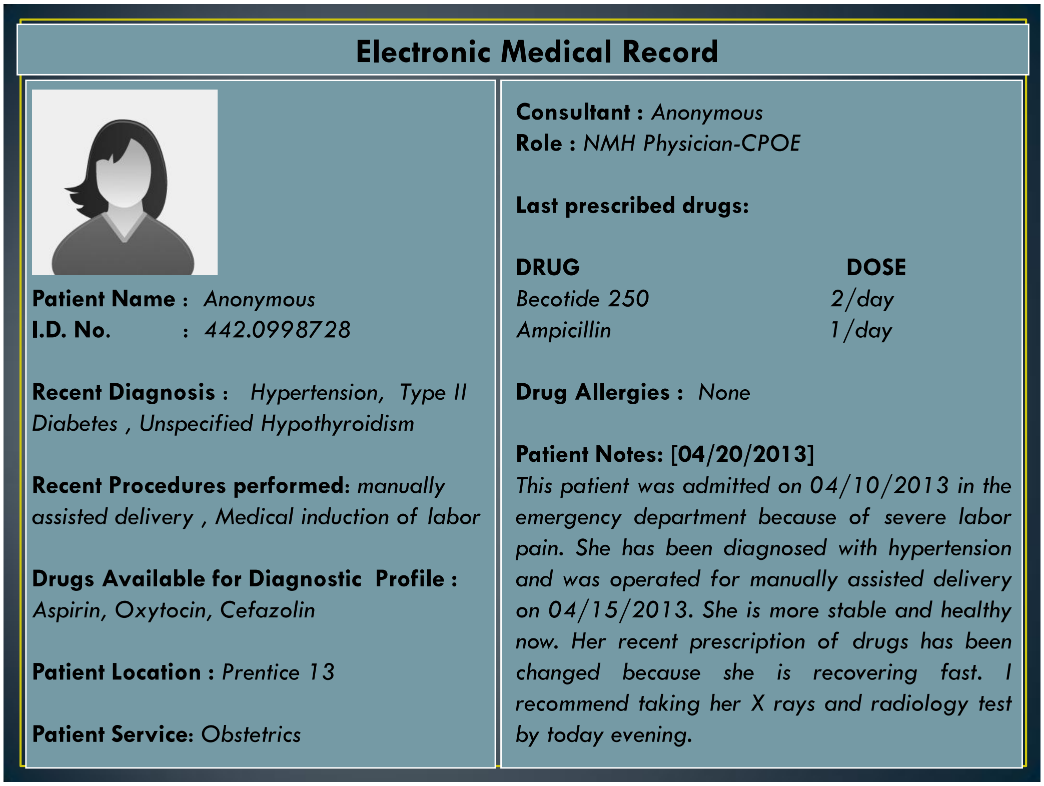
Introduction and Overview

Healthcare Information Technology is suffering because of lack of relevant data mining and information retrieval techniques to assist patient care. While many factors have been implicated in causing this issue, a lack of proper security measures looms largest. Despite the recent legislative efforts such as the HIPAA privacy rule [2] to legally mandate the protection of private healthcare information, issues still persist regarding patient privacy. These problems do not stem from the lack of awareness, but rather a lack of effective practical solutions. This paper aims to address one of the distinct challenges in implementing effective healthcare security measures: detecting anomalous activity. This central challenge, as well as some of the other mitigating problems plaguing the industry, is discussed in further detail in 1.2. However, to first provide a solid contextual understanding of the problem statement, an introduction into the basics of healthcare workflow dynamics and information architecture is provided in section 1.1. The section 1.2 refines the initial motivation and focuses the analysis on a specific component of the issues in healthcare security. A novel solution to this specific issue is described in section 1.3 and section 1.4 outlines the organization of this logic. [1]

1.1 Overview of Electronic Medical Record

The authors in [4] define EMR as an application environment composed of the clinical data repository, clinical decision support, controlled medical vocabulary, order entry, computerized provider order entry, pharmacy, and clinical documentation applications. This environment is used by many healthcare practitioners to document and monitor patient's electronic medical record

[1] The key contribution of this thesis will appear in "Proceedings of IEEE International Conference on Intelligence and Security Informatics, 2013" [3].



Electronic Medical Record

Patient Name : *Anonymous*
I.D. No. : *442.0998728*

Recent Diagnosis : *Hypertension, Type II Diabetes , Unspecified Hypothyroidism*

Recent Procedures performed: *manually assisted delivery , Medical induction of labor*

Drugs Available for Diagnostic Profile : *Aspirin, Oxytocin, Cefazolin*

Patient Location : *Prentice 13*

Patient Service: *Obstetrics*

Consultant : *Anonymous*
Role : *NMH Physician-CPOE*

Last prescribed drugs:

DRUG	**DOSE**
Becotide 250	*2/day*
Ampicillin	*1/day*

Drug Allergies : *None*

Patient Notes: [04/20/2013]
This patient was admitted on 04/10/2013 in the emergency department because of severe labor pain. She has been diagnosed with hypertension and was operated for manually assisted delivery on 04/15/2013. She is more stable and healthy now. Her recent prescription of drugs has been changed because she is recovering fast. I recommend taking her X rays and radiology test by today evening.

Figure 1.1: Mock EMR of a Patient

which may consist of diagnosis of the patient in an encounter, procedures operated on him, medication history, his location and services during the hospital stay etc. One main advantage of such a system is an ease of communication between the clinicians in different departments (i.e., nurses, physical therapists, and respiratory therapists) in the hospital for managing effective patient care. However, [4] further differentiates EHR from EMR. An EHR on the other hand is a subset of each care delivery organizations EMR, presently assumed to be summaries like ASTM's Continuity of Care Record (CCR) or HL7's Continuity of Care Document (CCD), is owned by the patient and has patient input and access that spans episodes of care across multiple CDOs within a community, region, or state (or in some countries, the entire country). An EMR is also used to provide documentation that a patient was seen or a test was performed so that the clinician can obtain reimbursement by an insurance company or government agency [5].

Figure 1.1 shows a sample EMR of a patient in any hospital. The Figure shows the medical record of an anonymous patient in the hospital and her health history. The record shows the patient was admitted on 04/10/2013 under severe labor pain. She was diagnosed with hypertension and later went through manually assisted delivery. The contents of the notes are taken by different clinicians accessing the EMR for patient care. This medical record may store very confidential information about a patient. There are several clinicians constantly accessing these EMR systems which give rise to many privacy concerns related to patient data being compromised. The authors in [6] discuss similar privacy and security considerations of medical records from being exploited by external harmful entities.

1.2 Problem Statement

Though many systems claim to implement role based access control but because of the dynamic nature of healthcare environment, it is more likely that these roles will break their access privileges in the case of emergency situations. In such cases, there is a potential threat to the patient information if the access is made because of some prior hidden motive. The lack of experience based access control [7] is one of the major security drawbacks in such systems. The condition is going to be worse when we have genomic data available in each patient's EMR. Imagine some curious users trying to access an EMR of a celebrity, an employer accessing EMR of an employee he wants to hire to assess his personality or even a girl seeking an EMR of the boy to approve marriage. Just by looking at the EMR anyone will be able to confirm the patient's health prospects, intelligence and family health history. Lack of publicly available data constraints the researchers to carry out any consequential research in the field of healthcare. In the hospital environment, access control is almost completely dominated by usability. Under no circumstances should an employee be denied access to resources that are critical to patient health. Also, the work load and the employment costs of health specialists are so large that it is economically inefficient to install time-consuming security measures. Therefore, the enforcement of access control policies is often very weak. To achieve compliance, all accesses to resources are logged in audit logs such that i) fraud can be detected in the future and ii) knowing that access is being tracked, employees comply in the presence. The main problem with this approach is that the sheer amount of data that is being saved exceeds what a human can manually overlook. This situation demands for automatic or semi-automatic approaches to analyze audit logs.

1.2.1 Summary of the Main Challenges

Let us summarize our discussion in 1.1 and 1.2 pertaining to the existing limitations of the security measures in the current EMR systems and lack of relevant mining techniques for automated patient care. Some of the major challenges in the present systems are as follows:

1. The current systems are based on Role Based Access Control where access is granted based on the roles of the clinicians accessing the medical record. But, healthcare is a dynamic and collaborative environment where the role privileges might need to be changed in the case of emergency situations. These systems are unable to detect unexplained accesses made by the clinicians in such an environment.

2. There are millions of access logs which get generated by the clinicians accessing patient medical records in short amount of time. There is a need to build high end auditing tools to analyze these huge access log dataset to learn and infer the characteristics of the clinicians based on their accesses to the patients.

3. There is no inference on the patient and clinician relationship based on the accesses made by these entities to the patients throughout the encounter. There is lack of data mining in place to mine such information and use it for patient care and security.

4. Current research in the field of EMR security does not focus on detection methods pertaining to flag clinicians having non-random but suspicious behavior in their access patterns. For instance, clinicians who are not accessing random patients, but have prior knowledge of the hospital workflow and interested in particular type of patients.

1.3 Approach to the Problem

Our approach to this problem is an anomaly detection framework that takes a history of audit logs as an input and builds a semantic relationship between the user-patient accesses by *typing* the user based on different kind of patients accessed by him. Then we can detect different types of anomalous users in the system. At the heart of the method (and at the heart of the problem really) lies the notion of typicality that must be established based on the input.

Much of the current work on anomaly detection in hospitals has focused on what one might call the *Random Object Access* (ROA) where the object in question is a patient electronic medical record. In this model, illegitimate users are modeled as ones who access patient records randomly. This is a plausible model since illegitimate accesses may look random as they are often based on features that lie outside the medical system (like accessing the record of a famous actor with an ordinary disease). However, the model may not apply well to the open terminal case since the users are not making random accesses; they are making accesses that are appropriate to their roles while ascribing these to a user for whom the action is not appropriate. Consider also a related problem, where a user is doing tasks that should be done by another user. This behavior may be inappropriate, but it will probably not look like accessing random patient records.

A key contribution of this thesis (to appear in [3]) is the following: This thesis proposes a new approach to anomaly detection 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 natural language processing, 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. In our case we apply these techniques to patients, who play a role similar to articles, whereas the data in the patient medical records plays the role of words. The idea is to derive a model of the topics in the hospital (common groups of properties of patients) and to use these to characterize the interests of users, who can be viewed as "readers" of these topics. An RTA user is one who accesses a collection of topics at random. This is subtly different from an ROA user, whose access to *patients* is random. The RTA is useful for detecting anomalous users who are systematic in using hospital topics but are potentially unusual in the combination of topics they access. Think of 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 obstetrics records. This may not be illegitimate but it might be appropriately flagged as anomalous.

Random Topic Access Detection (RTAD) is based on using Latent Dirichlet Allocation (LDA) to define topics and a k-nearest neighbor (k-NN) algorithm to detect RTA. The principle novelty of the work is not the use of these specific techniques for anomaly detection (since similar techniques have been used in other contexts); 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 the tendency to select many topics. This approach has advantages of generality over trying to simulate open terminals, masquerading, or other attack modes directly. LDA was ran on four months of patient records and created a list of hospital topics that enable each patient to be described as a mixture of topics. From this topics were derived that characterize users that access these patients and the hospital-assigned roles of these users. This work focuses on five roles and three different kinds of users, allocated between the two extreme cases where users favor a few topics strongly to users who favor many topics weakly. For each of these cases we computed the Area Under Curve (AUC) from Reciever Operating Curves (ROCs) for detecting RTA users. These studies have led to variety of findings. For instance, the effectiveness of RTAD varies based on the class of topics used (for instance, diagnoses or medications or both). Given the best approach topic model for each of the five roles, we find that RTAD works better for each role than it works on the collection of all users if they are described by a few topics, but performance declines for users that are more topic-agnostic in four of the five roles. More results of these kinds appear in the analysis section.

1.3.1 Summary of the Main Contributions

Now, let us summarize our discussion in section 1.3 to provide potential solutions to the challenges mentioned in section 1.2 and focus on the main contributions of this work. Some of the major contributions are as follows:

1. This work introduces and studies the random topic access model, RTA, aimed at the users whose access may well be illegitimate but is not fully random because it is focused on common hospital themes. For instance, gynecologist who suddenly is interested in the medical records of male patients diagnosed with erectile anxiety and depression.

2. We also propose RTAD framework for evaluating the ability to recognize various types of random users i.e. (Directed user, Random User, Indirect User). More information about these user types can be seen in the section 5.2.

3. Another important aspect of this work is the notion of User Typing to mine the given EMR information and audit log data to find useful patterns of clinicians accesses to the patient's EMR. This enables to characterize the given clinician and build a semantic relationship with the patient attributes such as diagnosis, procedures, medications, location.

4. We also use multidimensional scaling to visualize the clinicians and their probability distribution over patient attributes (derived from User Typing) on a 3D scale. This helps us to understand the distribution of users within the same role based on their actions. The users within a same role are expected to behave in a similar way.

5. Our method also provides a way to dynamically combine roles based on the similarity of the user distribution and the distance among different roles. We can cluster roles together based on the type of patients they access according to each attribute of the patient or their combination.

1.4 Structure of This Thesis

Chapter 2 describes in more detail the current research in Health Information Technology and gives thorough background information on the insider threat detection models for hospitals. Section 2.1 explains various clinical terminologies used in the EMR dataset such as ICD-9 Diagnosis, ICD-9-CM Procedures and RxNorm for medications. This helps us understand the different features in the given dataset. Sections 2.2, 2.3 and 2.4 explain various threat detection models.

Chapter 3 aims at providing insight into the Northwestern Memorial Hospital EMR and Audit log dataset. The data description in section 3.1 explains the architecture of the EMR and Audit log dataset respectively. Also, the relationship between both the dataset can be understood with respect to the diagnosis, procedures, medication, locations and services of the patient. Table 3.3 shows the 10 most frequent occurring features in the dataset.

Chapter 4 is the highlight of this thesis work and lays a strong foundation for the RTAD model discussed in Chapter 5. Section 4.1 introduces LDA Topic Model and explains it in the context to our problem. The section 4.3 formalizes the User Typing approach to derive user attributes from the patients and characterize the users in terms of patients. The topics derived in the section 4.1 are summarized in the section 4.3.

Chapter 5 formalizes the *Random Topic Access model* (RTAD) to detect various types of intruders in the system namely (direct, completely random and indirect users) in section 5.2, whereas section 5.1 explains the mathematical model behind k-nearest neighbor to detect outliers in the system.

Chapter 6 discussed the experiment and analysis of RTAD model discussed in Chapter 5, for anomaly detection for five major hospital roles. We compare AUC's across all roles and all values of alphas.

Chapter 7 concludes the thesis with summarization of results and potential future scope of this work.

Chapter 2

Background and Related Work

This chapter highlights the main contributions in the literature for insider threat detection for hospitals. The section 2.1 introduces various standard clinical terminologies used in most of the EMR systems to denote patients diseases, surgeries and medications. Section 2.2, 2.3 and 2.4 we highlight the main differences between our approach and the existing models.

2.1 Clinical Terminologies

Medicines complex language is represented in clinical terminology and vocabulary systems. Clinical terminologies represent terms related to the medical field while vocabularies are collections of terms. Both clinical terminologies and vocabularies provide a way to capture detailed data in an electronic health record (EHR). They support the transformation of paper-based to electronic records by providing a machine-readable data structure. Clinical terminologies are considered the input format while classification systems are the output format. Three main terminologies used in the analysis are ICD-9 Diagnosis [8], ICD-9-CM [9] Procedures and RxNorm [10].

2.1.1 ICD-9 Diagnosis

The International Classification of Diseases, Ninth Revision, (ICD-9) [8] is based on the World Health Organizations Ninth Revision, International Classification of Diseases (ICD-9). ICD-9 is the official system of assigning codes to diagnoses and procedures associated with hospital utilization in the United States. ICD-9 is used to code and classify mortality data from death certificates. ICD-9 consists of a numerical list of the disease code numbers in tabular form; An alphabetical index to the disease entries, and A classification system for surgical, diagnostic, and therapeutic procedures . ICD-9 diagnosis codes can be categorized into three levels. The lowest level consists of thousands of diagnosis chapters, where each chapter is a particular diagnosis code. The middle level has about 1000 diagnosis chapters where each diagnosis chapter might contain group of codes based on similarity of diagnosis. The highest level consists of 19 Diagnosis chapters binded together as a high level concept of different diagnosis. The structure of these chapters are represented as trees. For example a diagnosis 500 can be divided further but can be mapped directly to 500(Highest level) or 501(Middle Level) or 501.2(Lowest Level). The values are in the range [0, 1000], E[0-999], V[0-99].

For instance: ICD-9 Diagnosis codes for "complications of pregnancy, childbirth, and the puerperium" are in the range [630-679]. Hence all different categories of related diagnosis are listed in the same range. They can be further divided into seven different categories i.e. [630-633], [634-639], [640-649], [650-659], [660-669], [670-676] and [678-679]. For example, range [660-669] is classified as "Complications occurring mainly in the course of labor and delivery" where 660 is 'Obstructed labor' and can be further divided into four subcategories- i) [660.01]: Obstruction, malposition, delivered ii) [660.11]: Obstruction, bony pelvis, delivered iii) [660.41]: Shoulder dystocia, delivered iv) [660.61]: Trial of labor, failed, delivered.

2.1.2 ICD-9-CM Procedures

ICD-9-CM procedure codes [9] are based on the official version of the World Health Organizations Ninth Revision, International Classification of Diseases, Clinical Modification (ICD-9-CM) contains codes for operations and procedures performed on an inpatient basis. The range varies from [0,100]. The highest level consists of 18 Procedure chapters binded together as a high level concept of different procedure. The structure of these chapters are represented as trees.

For instance: ICD-9-CM Procedure codes for "Obstetrical Procedures" are in the range [72-75]. Hence all different categories of related procedures are listed in the same range. They can be further divided into four different categories i.e. [72.0-72.9], [73.01-73.99], [74.0-74.99] and [75.0-75.99]. For example, range [74.0-74.99] is classified as "Cesarean section and removal of fetus", which can be further divided into six subcategories- i) [74.0]: Classical cesarean section ii) [74.1]: Low cervical cesarean section iii) [74.2]: Extraperitoneal cesarean section iv) [74.3]: Removal of extratubal ectopic pregnancy v) [74.4]: Cesarean section of other specified type vi) [74.91-74.99]: Cesarean section of unspecified type.

2.1.3 RxNorm

RxNorm [10] is two things: a normalized naming system for generic and branded drugs; and a tool for supporting semantic interoperation between drug terminologies and pharmacy knowledge base systems. The National Library of Medicine (NLM) produces RxNorm. Hospitals, pharmacies, and other organizations use computer systems to record and process drug information. Because these systems use many different sets of drug names, it can be difficult for one system to communicate with another. To address this challenge, RxNorm provides normalized names and unique identifiers for medicines and drugs. The goal of RxNorm is to allow computer systems to communicate drug-related information efficiently and unambiguously.

For instance: RxNorm group source data into collections of synonyms (called concepts). The drug Naproxen can be referred as Naproxen Tab 250 MG, Naproxen 250mg tablet (product), NAPROXEN @250 mg@ORAL @TABLET, Naproxen 250 MILLIGRAM In 1 TABLET ORAL TABLET, NAPROXEN 250MG TAB,UD [VA Product]. Although the drug names in this Naproxen example appear different, they all have the same meaning at a certain level of abstraction. RxNorm groups these as synonyms into one concept by various relationships like - "ingredient of", "isa", "has dose form" or "has ingredient".

2.2 Random Object Access Model

The ideal way to evaluate the effectiveness of an intrusion detection system is with actual intruders. This is quite difficult for studies in the area of hospital information systems since there is very little real data of this kind available for study. The notable exception is a study of supervised learning [11] where real intrusion data and accesses deemed suspicious by administrators were used to train a classifier that achieved reasonable results. Getting this type of data and access to these sorts of experts is very difficult. Moreover, the ability of “expert” analysts is somewhat questionable, since most hospital personnel are not trained at detecting security violations.

For these reasons, most work 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. For instance, MetaCADS [12] does anomaly detection based on Singular Value Decompositions (SVD) of user-patient access and patient-feature matrices. Their evaluation centers on the recovery of users generated via ROA using k-NN with eigenvectors weighted by their eigenvalues. This approach obtains good AUC values for ROA. Our analysis with RTAD considers subpopulations of users by role based on RTA. Additionally, although MetaCADS utilizes SVD for dimensionality reduction, the resulting features only group dimensions together according to their contribution to the variance to the system, not their semantic coherence.

It might be useful to note that MetaCADS was designed to find anomalous users whereas SNAD [13] was designed to find specific anomalous accesses of the users. EBAS [14] determines reasons for the access to patients by the users based on the assumption that for each department the employees are responsible for specific diagnosis. The anomaly detection framework works on the level of hospital-designated departments rather than individual user probability for the given diagnosis. The assumption is that departments are groups of employees with similar responsibilities who behave similarly. 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.

There is some literature about detecting anomalous records in more generalized dataset. For instance, [15] uses association rule mining and Bayesian approaches to discover outliers in categorical dataset. This work does not account for heterogeneous dataset such as electronic medical record (EMR) systems and the associated user-level access logs. It also assumes no prior information regarding the interaction of attribute sets, something which could be a valuable resource for healthcare professionals.

2.3 Masquerading User Model

There is also a good body of work [16, 17, 18, 19] 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 activities that are a deviation from expected activities. These techniques provide an interesting contrast with the ones we present below. Authors in [19] have tried to create threat detection models for masquerading users in GUI based systems by extracting relevant features through SVD and using supervised learning to classify anomalous behavior. Our study differs from their approach in two respects: the first is in the size of the dataset. While [19] only includes the activity of three different users, we have annotated information for thousands of users and patients collected over a four year span. The second is with respect to the feature space. We argue, once again, that the latent topic space provides more semantically coherent and intuitive features than those obtained through SVD. With respect to [18], they use a Markov Model trained on real user commands and then show that if they generate commands using the Markov Model and inject these commands into classifier, they are not detected. Even though the commands were not generated by a user. It is not anomaly detection but a way of circumventing anomaly detection. The authors in [16] and [17] extend research in masquerade detection using UNIX commands issued by the users and applied supervised learning mechanism to predict the anomalous users but these approaches generally do not scale well for large access logs in a dynamic environment such as healthcare.

2.4 Process Mining Model

Process mining is useful for at least two reasons. First of all, it could be used as a tool to understand how people and/or procedures really work. Process mining could be used to gain insight into the actual process, e.g., the flow of patients in a hospital. In such an environment, all activities are logged, but information about the underlying process is typically missing. Secondly, process mining could be used for comparing the actual process with some predefined process model. Such a model specifies how people and organizations are expected to work. By comparing the descriptive or prescriptive process model with the discovered model, discrepancies between both can be detected and used to improve the process and identify deviations and anomalies. [20] and [21] illustrates the use of heuristic mining and fuzzy mining to understand the hospital workflows. In general, heuristic mining is suitable for accounting the errors, deviations, and random activities in real world events. Fuzzy mining provides a way to handle complex and unstructured process by focusing on important activities and relations.

Chapter 3

Dataset Analysis

Dataset analysis is an important step in heading towards formalizing our RTAD model. The dataset from Northwestern Memorial Hospital are central to the thesis study and have been collected by Cerner Systems over a period of four months. It consists of over 5 million accesses made by 8000 clinicians to 14,000 patients and has two main parts i) Electronic Medical Record data ii) Audit Log data. The EMR dataset contains patient's health history including diagnosis, procedures and medications in an encounter, whereas the audit log data contains the information about the user and his accesses to the patients at a given time and patient - location and service. Section 3.1 describes the entire dataset in detail. In section 3.2 we analyze some important statistics about the dataset, which will help us understand the data.

3.1 Data Description

Our dataset consists of all user accesses, or audit logs, made over a four month period, in addition to Electronic Medical Record (EMR) data for patients admitted in this time frame. Patient and user IDs were anonamized for security purposes. As a noise reduction measure, we filtered all out-patient entries, focusing only on patients who stayed a significant amount of time at the hospital (more than 24 hours). We further removed all 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. After preprocessing, our 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. In terms of the information characterizing users, audit logs provide details regarding the service and location of a given user accesses; EMR data, on the other hand, provides information regarding the diagnosis, medication, and procedures ordered for a particular patient. Service, location, diagnosis, medication, service/location, and procedure are referred to as dimensions in our analysis. The following subsection aims to elucidate the key differences between these two distinct datasets.

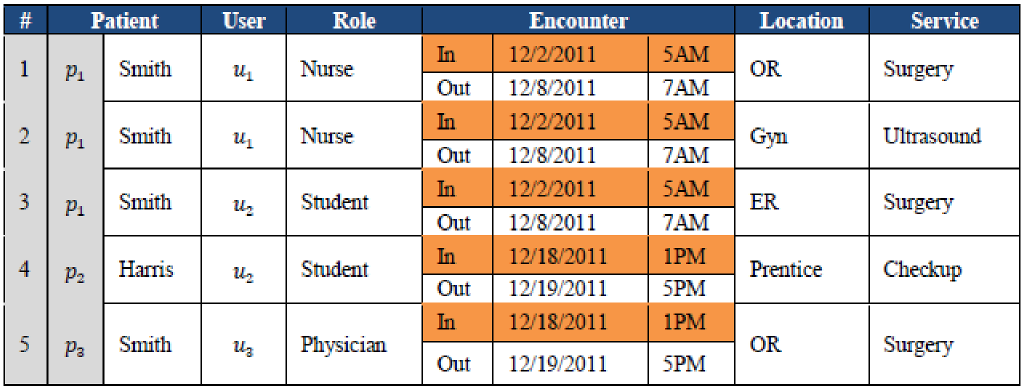
3.1.1 Audit Log Data

Audit logs consist of user accesses to specific patients, logging the service the user provided and the location of the access. Service can assume one of 30 different values while location can assume one of 49 different values. Table 3.1 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.

Figure 3.1 consists of 4.9 million accesses where each access is represented in the form of an access. An access consists of a unique patient P and user U having role R, the encounter information signifies the period of stay of the patient in the hospital. Each access has multiple attributes for the patient and the user. For instance, patient was accessed when he was at location L and for service S. In the table we can see that patient Smith has been accessed multiple times by different users at different locations and services. The same user also accesses other patients. We can imagine this as a bipartite graph where all the edges are from user and patient. Each user has a unique role in the given dataset and a role can have multiple users.

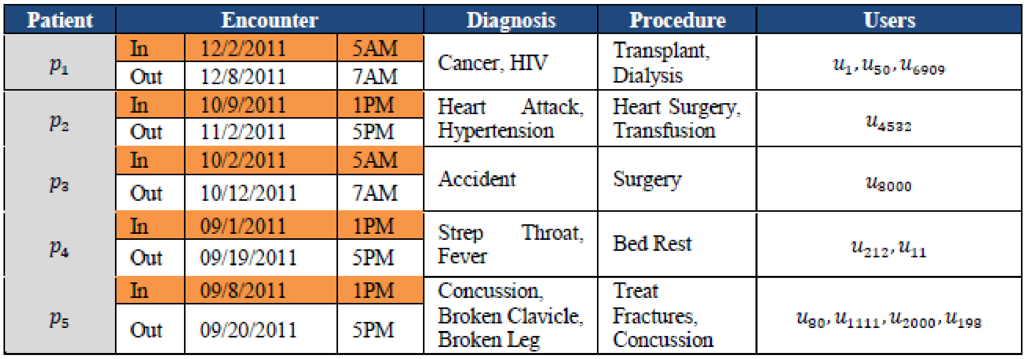
3.1.2 EMR Data

EMR data in Figure 3.2 consist of different patient records, with each record corresponding to various diagnosis, procedures, and medications. A given dimension is a binary vector, with each bit in the vector representing the



#	Patient		User	Role	Encounter			Location	Service
1	p_1	Smith	u_1	Nurse	In	12/2/2011	5AM	OR	Surgery
					Out	12/8/2011	7AM		
2	p_1	Smith	u_1	Nurse	In	12/2/2011	5AM	Gyn	Ultrasound
					Out	12/8/2011	7AM		
3	p_1	Smith	u_2	Student	In	12/2/2011	5AM	ER	Surgery
					Out	12/8/2011	7AM		
4	p_2	Harris	u_2	Student	In	12/18/2011	1PM	Prentice	Checkup
					Out	12/19/2011	5PM		
5	p_3	Smith	u_3	Physician	In	12/18/2011	1PM	OR	Surgery
					Out	12/19/2011	5PM		

Figure 3.1: Mock Log Table of User Accesses for Service and Location



Patient	Encounter			Diagnosis	Procedure	Users
p_1	In	12/2/2011	5AM	Cancer, HIV	Transplant, Dialysis	u_1, u_{50}, u_{6909}
	Out	12/8/2011	7AM			
p_2	In	10/9/2011	1PM	Heart Attack, Hypertension	Heart Surgery, Transfusion	u_{4532}
	Out	11/2/2011	5PM			
p_3	In	10/2/2011	5AM	Accident	Surgery	u_{8000}
	Out	10/12/2011	7AM			
p_4	In	09/1/2011	1PM	Strep Throat, Fever	Bed Rest	u_{212}, u_{11}
	Out	09/19/2011	5PM			
p_5	In	09/8/2011	1PM	Concussion, Broken Clavicle, Broken Leg	Treat Fractures, Concussion	$u_{80}, u_{1111}, u_{2000}, u_{198}$
	Out	09/20/2011	5PM			

Figure 3.2: Mock EMR Table of Patients for Diagnosis and Procedure

presence or absence of a particular feature. A feature in this case is dependent upon the dimension. Diagnosis, for instance, is characterized by the lowest level of the ICD-9 code hierarchy, with 4543 unique codes [8]. Procedure is similarly defined with respect to ICD-9-CM, albeit with less code words (1237). Medication features is defined with respect to RxNorm, a normalized naming system for generic and branded drugs [10]; in total, there are 642 codes. Table 3.2 provides basic statistics on the various dimensions.

It is important to note the lack of precise 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. This is the actual medical record of the patient which is accessed by the users. The main difference between the access log and EMR is that the access log is the summary of accesses to the EMR. The data in the patients medical record

Table 3.1: Summarized NMH Audit Log Statistics

Attribute	**Value**
Duration of Audit Logs	4 months
Distinct Accesses	4,979,465
Distinct Patients	12,488
Distinct Patient-Encounters	14,606
Distinct Users	7,932
Average Patients Accessed per User	115
Average Accesses of Patient	340

consists of his personal information (address, name, and identity) as well as health information which indicates the history of his procedures, medications and diagnosis throughout the stay of the hospital. A patient can visit hospital multiple times. If a patient visits a hospital and is not admitted, then is considered as outpatient. If a patient is admitted for more than 24 hours, he is considered as inpatient. For instance, in the table 3.2 we can see that the patient P during the stay in the hospital has been diagnosed with cancer and HIV and gone through multiple procedures i.e transplant and dialysis. In the access log table, the data consists of the number of times $user_1$, $user_{50}$, $user_{6909}$ accesses patient P_1 record in the given encounter period.

3.2 Data Statistics

The NMH dataset is rich with different type of patient and user information and needs to be summarized in order to understand it. Table 3.1 and 3.2 summarize the basic data statistics from the NMH Audit Logs and EMR. Also, table 3.3 shows the 10 most frequent diagnosis, procedures, medications, service and locations in the patient population. The statistics show that 27% of the patient population has been diagnosed with hypertension, 74.8% have been given aspirin in some form, 27.5% have gone through fetal monitoring procedure, 26.6% of the patients are on obstetrics service and 28.7% have been through the location Prentice 8 Labor & Delivery. These statistics shows a very high degree of confidence that the dataset is skewed towards the obstetrics population. On further analysis we find that 66.7%

Table 3.2: Summarized NMH User-Patient Statistics

Attribute	**Value**
Distinct User Roles	156
Distinct Patient locations	49
Distinct Patient services	30
Distinct Patient diagnoses	4,543
Distinct Patient procedures	1,237
Distinct patient medications	996

of the patients are female and most of them are in the age between 20 and 40. Figure 3.3 summarizes the access frequency patterns of the users and roles in the NMH access log dataset. We observe that the access pattern of the users are highly skewed. The highest accesses made by any user is 17,054 out of 4.9 million accesses. 21.3% of the users have more than 1,000 accesses in total in the audit logs. Figure 3.3 also shows the distribution of role accesses in the NMH access log dataset. We find that most of the roles have very high access frequency while others have very few accesses made in the four month time period. Patient Care staff nurse made the maximum number of accesses which amounts to 37.5% of the total accesses, followed by NMH Resident/Fellow-CPOE, Patient Care Assistive Staff, Patient Care Staff Nurse (Pilot), NMH Physician-CPOE etc. 50.6% of the roles have less than 1,000 accesses which shows the sparsity in the access of atleast 80 roles and the users within those roles. Figure 3.3 also shows the diversity of access patterns for users to patients. The maximum unique patients accesses by a particular user is 20.1% of the total patient population. 0.7% of the users access more than 1,000 patients and 6.3% of the users access only 1 patient. Finally, in Figure 3.3 we analyze the number of users within each role. This graph analyzes the sparsity between the roles with respect to the number of users within each role. One unique property of the dataset is that each user has only one role and each role can have multiple users. Our analysis show that 18% of the users have the role Patient Care Staff Nurse, 8.9% of the users are NMH Resident/Fellow-CPOE, 8.3% of the users are NMH Physician-CPOE, 5.9% of the users are Med Student-CPOE, 5.5% of the users are Physician Office. We also observe that 14.5% of the roles have only one user. These statistics help us to gain an insight into the data distribution

Table 3.3: NMH summarized EMR: Top 10 for each Dimension

Dimension	Top 10
Diagnosis	Hypertension, Outcome of delivery, Other and unspecified hyperlipidemia, Unspecified anemia, Type II or unspecified type diabetes mellitus without mention of complication, Second-degree perineal laceration, with delivery, Esophageal reflux, Coronary atherosclerosis of native coronary artery, Other current maternal conditions classifiable elsewhere with delivery, Unspecified hypothyroidism
Procedure	Other fetal monitoring, Manually assisted delivery, Repair of other current obstetric laceration, Other artificial rupture of membranes, Low cervical cesarean section, Medical induction of labor, Puncture of vessel, Transfusion of packed cells, Hemodialysis
Medication	Aspirin, Docusate, Bupivacaine, Potassiumchloride, Glucose, Esomeprazole, Lactated Ringers Injection intravenous solution, Dalteparin, Ondansetron, Cefazolin
Service	Obstetrics, Hospital Medicine, General Medicine, Orthopedics, Gynecology, Neurosurgury, Cardiology, Transplant Surgury, Hematology, Urology
Location	Prentice 8 Labor & Delivery, ASU Recovery 65, Emergency Department 1, Prentice 1 OB Triage, Prentice 11, Prentice 13, Prentice 12, Fienberg 15 E, Fienberg 16 W, Fienberg 10 W

across all users and roles. Now we will highlight the most frequent attributes in the dataset.

Table 3.3 shows 10 most frequent occurring patient dimensions in the summarized NMH EMR dataset. The dataset is mostly skewed towards obstetrics patients, which is also evident from table 3.3. Almost all the dimensions and their top features concur with obstetrics. Majority of the patients have been diagnosed with Hypertension and delivery which are also highly correlated in the EMR dataset. The intuition is that the majority of the females in the hospital who are diagnosed with both these problems go through correlated procedures, service and location in the hospital. Other patients are diagnosed with cardiovascular, diabetics etc, which also explains another category of patients having heart problems. Similarly, most of these patients who have been diagnosed with delivery, go through obstetrics procedure such as manually assisted delivery, fetal monitoring, repair of the current obstetric laceration, Low cervical cesarean section etc. The highly correlated patient location in the hospital with such diagnosis and procedure is Prentice 8 Labor & Delivery, Prentice 13, Prentice 11 and Prentice 12. The hospital location ASU Recovery is a place where patients go after a surgery. Hence, majority of the patients who go through a procedure, also go through ASU Recovery. Other locations include emergency department and Fienberg which handle all kind of patients. Majority of the patients are given aspirin to relieve minor aches and pains. Others are on different types of drugs such as Esomeprazole, Dalteparin, Cefazolin etc.

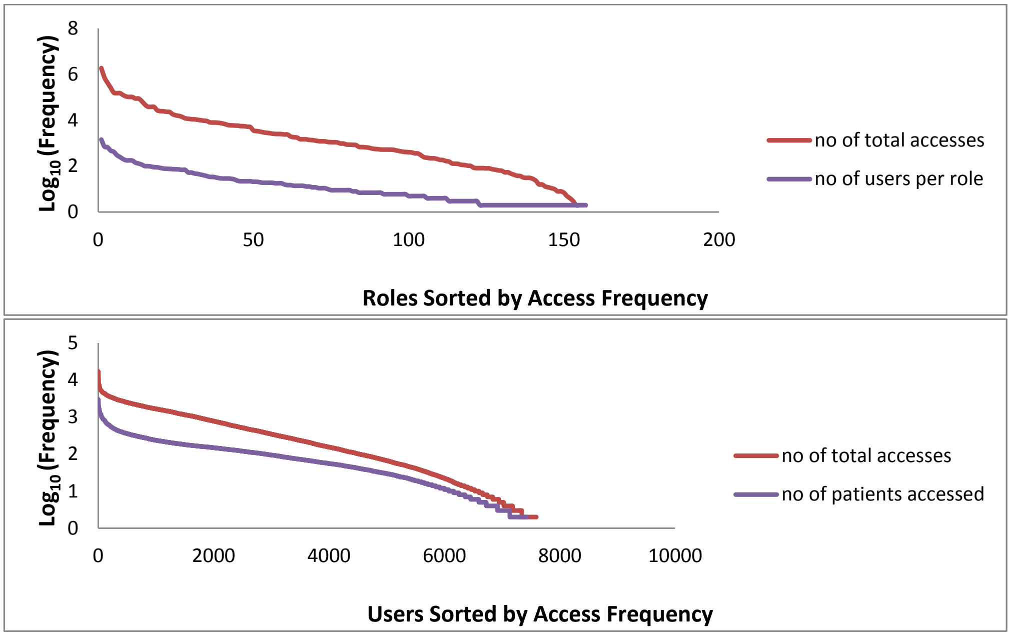


Figure 3.3: NMH Audit Log Access Summarization

Chapter 4

Topic Modeling

The dataset described in Chapter 3 is high dimensional data for all attributes of patients which makes it essential to uncover the underlying semantic structure of the features. Latent Dirichilet Allocation (LDA) [1] is the most promising approach of dimensionality reduction which is a class of probabilistic topic models. The intuition behind LDA is that a document contains mixture of different topics/themes. In our case it is analogous to the fact that patient can be represented over distribution of topic/themes for each high dimensional attribute (say diagnosis, procedure or medication). For instance, in case of diagnosis, the patient can be described as a probability distribution over topics of diagnosis. Each topic of diagnosis consists of probability distribution over different ICD-9 diagnosis codes, which makes the semantic structure of the topic intuitive and can be analyzed manually or peer review.

For example, consider using themes to explore the complete history of the patients with mental diseases in the hospital. At a broad level some of the themes in the hospital might correspond to the neuro, cardio, obstetrics etc. We could zoom in on a theme of interest, such as mental illness, to reveal various aspects of it such as alzhimer, brain tumour etc. We could then navigate through time to reveal how these specific themes have changed. And, in all of this exploration, we would be pointed to the original patients relevant to the themes. This chapter is organized as follows: In section 4.1 and 4.2, we will discuss the mathematical model behind LDA algorithm and how it can be used to reduce dimensions of the patients in our dataset. Section 4.3 formalizes the User Typing, which is one of the main contributions of this thesis. We define the user and patient relationship based on the topic derived in section 4.1 and use it to characterize the user based on the given dataset. In section 4.4, we summarize the topics derived in section 4.1 and give an intuition of the general themes in the hospital related to different dimension of the patient. In section 4.5 we use Multi Dimensional Scaling to visualize the users derived in section 4.3 to gain more intuition into the distribution of users within same role.

4.1 Latent Dirichilet Allocation

In this section we explain how we extract different features from the access logs and how to represent user actions such that different actions can be reliably discriminated from each other.

At the heart of our anomaly detection method is a semantic representation of the access log data given as an input. In order to detect illicit access to resources we must represent the data at hand in a way that (1) enables a numerical comparison between observations and (2) preserves semantic information about the medical aspects of what has been done in atomic user actions. Imagine how a domain expert would manually analyze a given audit log. Clearly, the main approach would be to relate the user actions with what makes sense from a medical perspective and from the processes that are typical in the respective hospital. Similarly, we pre-process the given data such that observed user actions are expressive with respect to the medical aspects and processes.

In an attempt to quantitatively represent the given audit logs and to fulfill the requirements given above, we employ Latent Dirichlet Allocation (LDA) [1]. LDA provides a set of topics, each represented as a bag of words that typically arise. For each patient, we will get an allocation in the topic simplex. We can then substitute the patient ID in each event with this vector. We then derive the topic distribution of the users accessing the patients by user typing in section 4.3 The dataset described above is high dimensional data for all attributes of patients which makes it essential to uncover the underlying semantic structure of the features. LDA is the most promising approach of dimensionality reduction which is a class of probabilistic topic models. The intuition behind LDA is that documents exhibit various topics. In our case it is analogous to the fact that patient can be represented over distribution of topics for each high dimensional attribute. For instance, in case of diagnosis, the patient can be described over topics of diagnosis. Each topic of diagnosis consists of probability distribution over diagnosis for that topic which makes the semantic structure of the topic intuitive and can be analyzed manually or peer reviewed.

From the perspective of our problem (we use diagnosis data as an example, but the same scheme holds for diagnosis and other features as well), LDA assumes the following generative process of patient data.

1. Choose $N \sim \text{Poisson}(\xi)$
2. Choose $\theta \sim \text{Dir}(\alpha)$
3. For each symptom $w_n,\ n \in \{1, \ldots, N\}$:
   (a) Choose a disease $z_n \sim \text{Multinomial}(\theta)$
   (b) Choose a symptom $w_n \sim p(w_n \big| z_n, \beta)$.

Here, $p(w_n \big| z_n, \beta)$ is a multinomial symptom distribution conditioned on disease z_n. In a more generic setting, we will refer to topics rather than diseases and words instead of symptoms. The graphical model of LDA is described in Figure 4.1. Filled circles represent observed random variables. Empty circles are latent random variables. Arrows indicate statistical dependencies. Entities on a plate with integer N exist in N different versions. In the case of diagnoses, θ indicates the probabilities for diseases, z_n indicates presence or absence of a disease for a patient out of M patients, and w indicates symptoms of a patient.

RTA entails modeling users as probability distributions over topics. These topics are defined with respect to the patient dimensions introduced in Chapter 3, supersets, or concatenations of these dimensions. We utilize a latent topic framework because modeling users as distributions over topics permits two important operations; the first is for users to be summarized in a semantically coherent way with respect to the entire user population. 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. Topic modeling, however, gives a concise description of not only how a user behaves in the context of his peers, but what the meaning of that behavior is. The second is it provides a mechanism for user generation. By modeling users as samples from a Dirichlet distribution over topic multinomials, a larger space is afforded of realizable users than the dataset may provide.

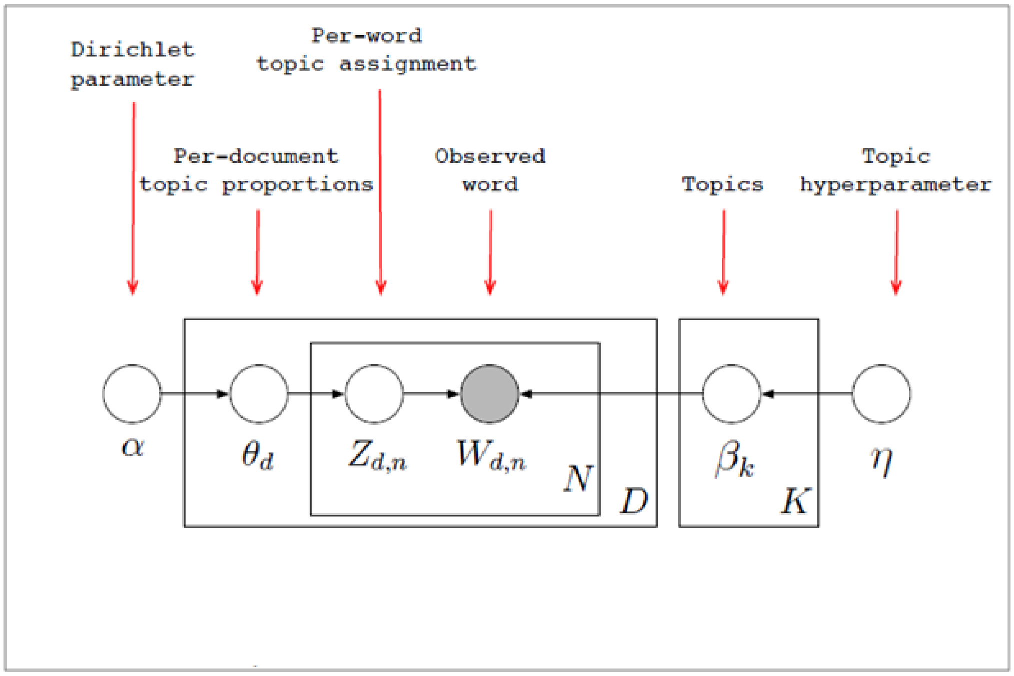


Figure 4.1: Graphical Model for Latent Dirichlet Allocation (LDA) [1].

This also has implications with respect to controlling the types of adversaries in different threat models, which will be elaborated on in Chapter 5.

Nevertheless, topic modeling provides a lot of power and flexibility in terms of characterizing and modeling users. While many algorithms exist for this purpose, the standard algorithm to use is Latent Dirichlet Allocation (LDA), which we use for our implementation RTAD. LDA is a generative model that attempts to model documents in a corpus as multinomials over a set of latent topics [1]; in turn, 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 get a particular topic distribution. This can be thought of as sampling from a $d-1$ simplex (probability space) controlled by a parameter to the Dirichlet: α. Typically denoted the concentration parameter, α controls where on the simplex multinomials are likely to be sampled. As $\alpha \to 0^+$, probability density is pushed towards the edges of the simplex, favoring multinomials heavily biased towards few topics. As $\alpha \to 1$, probability density becomes more evenly distributed around the simplex, making any multinomial equally probable of being selected. Finally, as $\alpha \rightarrow +\infty$, 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 apriori. We used α as a control parameter to generate users at five different levels of topic-concentration reaching from users whose actions are described by almost all equally contributing topics to users described by a single topic only.

In modeling users, we decided to generate topics for diagnosis, medication, procedure, and service/location. Additionally, we generated topics over the superset of all these dimensions, denoted mixed, and the concatenation of the topic vectors of these different dimensions, denoted combined. The logic behind modeling users in terms of mixed was to see whether or not heterogeneous combinations of dimensions are more informative than considering them in isolation. Likewise, we looked at combined to see if a naive concatenation of these different information types is comparable to or favorable to mixed. Section 4.4 provides a summarization for topic coherency across these different dimensions. Coherent topics are chosen for the service/location, procedure, diagnosis, and medication dimensions; the top 10 most probable words are displayed over 8 topics for each dimension. There is a strong bias in these distributions towards women's health, specifically child birth, demonstrating the efficacy and power of LDA to capture relevant semantic summarizations. It should be noted that with respect EMR derived dimensions, LDA was performed on patients. We will use the section 4.3 to reveal how these patient probabilities are translated in terms of users.

4.2 Perplexity Measure

Since the number of topics need to be chosen apriori, we utilized the perplexity measure, designed to assess the effectiveness of different topic numbers. The perplexity measure is an estimation of the expected number of equally likely words in the population; minimizing perplexity corresponds to maximizing the topic variance captured by the system [1]. We performed this analysis for each topic distribution, the number of topics corresponding to the minimum perplexity is shown in Table 4.1. Figure 4.2 shows the various perplexity values at different values of k (no. of topics) for procedure. To select the number of topics, we use the perplexity measure proposed in [1]:

$$\text{perplexity}(D_{\text{test}}) = \exp\left\{-\frac{\sum_{d=1}^{M}\log p(w_d)}{\sum_{d=1}^{M} N_d}\right\} \tag{4.1}$$

Here, D_{test} is a collection of patient data such as a collection of M diagnosis, w_d are the symptoms of the patients. Perplexity measures the data likelihood. High likelihood leads to a low perplexity measure. It is seen as a good measure of performance for LDA. We keep a holdout sample, train LDA on the rest of the data, and then calculate the perplexity of the holdout.

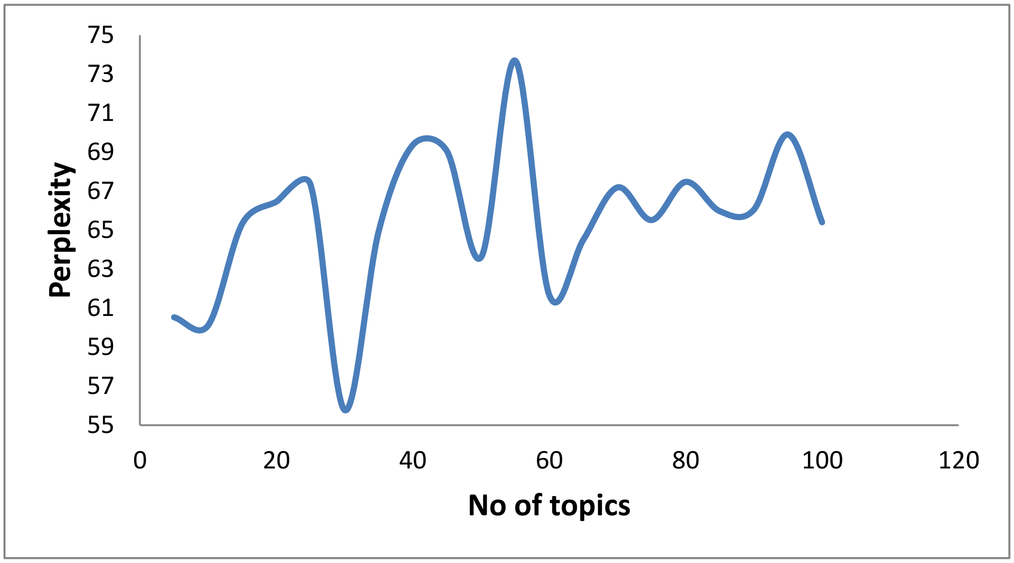


Figure 4.2: Perplexity Measure for Procedure

4.3 User Typing

User typing involves characterizing user behavior according to a particular dimension and data source. User typing with respect to Access Log data entails accumulating all accesses of a user and performing LDA, where each user is treated as a separate document. Typing with respect to Electronic Health Record data is much different. Because EHR data is aggregated over all users with respect to patients, no information exists regarding which set of records a particular user contributes to the patient. As such, in this dataset, users are characterized according to the types of patients they access, though

Table 4.1: Topic Summary

Dimension	# Of Topics
diagnosis	25
procedure	25
medication	25
service/location	20
mixed	40
combined	95

no assumptions are made regarding the extent to which a user contributes to the description of an accessed patient.

In the following sections, we aim to provide a more concrete mathematical notation for these definitions. In each section, suppose the dimensions in Access Log data is given by D_{AL} = Service, Location and EHR data is given by D_{EHR} = Diagnosis, Medication, Procedure. Furthermore, let U indicate the set of all users, P_u denote the set of all patients a user accesses, and A_u, the set of all access made by a user. The purpose of the following sections is to develop feature vectors for each user according to the two data sources.

4.3.1 User Typing With Access Log Data

For audit log data, users are summarized by LDA through the aggregation of their accesses and performing LDA on those aggregates. However, with respect to EHR data, 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 EHR dataset. Suppose each user accesses according to some dimension $d \in D_{AL}$ described by a vector $\vec{a_{u,d}}(i)$ where i indicates the i'th access of the user u. The raw feature vector, $\vec{f_{u,d}}$ for user u is simply the sum of all of u's accesses, A_u, shown in the following:

$$\vec{f_{u,d}} = \sum_{i \in A_u} \vec{a_{u,d}}(i) \tag{4.2}$$

The topic probability vector, $\phi_{u,d}$, is derived from transforming $\vec{f_{u,d}}$ via

LDA. Essentially, $\phi_{u,d}$ is a categorical probability distribution where for each $d_i \in d$, $\phi_{u,d_i} = P(d_i|u)$.

$$\phi_{u,d} = \text{LDA}(\vec{f_{u,d}}) \tag{4.3}$$

4.3.2 User Typing with EHR Data

As mentioned before, typing users with EHR data is slightly more complicated, due to the relatively limited amount of information about users in the EHR dataset. There are two possible approaches to typing in this manner. The first involves performing LDA on each patient individually and aggregating the topic probabilities, referred to as *post-aggregation.* The second involves aggregating raw patient vectors for each user and then performing LDA on the user, referred to as *pre-aggregation.* Pre-aggregation is very similar to user typing with Access Log data. Additionally, *post-aggregation* was used in the experimental procedure. Therefore, we will proceed with a discussion of the more complicated *post-aggregation.*

Suppose a patient p of user u has a record according to some dimension $d \in D_{EHR}$ described by a vector $\vec{f_{p,d}}$. The topic probability vector, $\phi_{p,d}$, much like in typing Access Log data, is simply the LDA transformation of $\vec{f_{p,d}}$. To derive the topic probability vector of the user, $\phi_{u,d}$, the topic probabilities of each of the patients u accesses calculated through a weighted averaged. In particular, let $N_u = |A(u)|$ denote the total number of accesses made by user u and $n_u(p) = |A_p(u)|$ refer to the number of times u has accessed p. The probability u selects p is simply $n_u(p)/N_u$. The weighted average of each of the patients topic probabilities is simply the expected topic probability distribution with respect to the probability of being selected.

$$\phi_{u,d} = E[\phi_{p,d}] = \sum_{p \in P(u)} \frac{n_u(p)}{N_u} \phi_{p,d} \tag{4.4}$$

This is permissible because the average of two categorical distributions is still a categorical distribution.

4.4 Topic Summarization

This section is based on the analysis done on the EMR dataset in the sections 4.1 and 4.2. In summary, LDA was run on the set of documents where each patient is treated as a document and the words in the documents are the features of the patient for a particular dimension (say diagnosis) collected from his/her electronic medical record. The input is also the number of topics, which is derived using perplexity measure in 4.2. The topics below can be seen as themes in the hospitals for given dimension of the patient. For instance, patient will have probability distribution of these topics/themes which will give us the bias of this patient towards a certain theme and to which he/she belongs. We picked 8 random topic distributions from each dimension and give an intuition behind the derived topics and their thematic structure for diagnosis, procedure, medications and service/location in Figures 4.3, 4.4, 4.5 and 4.6 respectively.

Summarization of Diagnosis Topics: In Figure 4.3, we observe that the *topic 04* and *topic 21* are highly related to women diagnosed with either maternal conditions or delivery. Patients having high probability of *topic 04* and *topic 21* are most likely women, diagnosed with obstetrics related problems. The *topic 01* is highly related to the patients diagnosed with irregularities in their blood (i.e. lack of white/red blood cells). Most of the words with high probability in this topic are related to blood problem. *Topic 17* on the other hand seems to be a mixture of different kind of diagnosis, mostly related to hypertension, also related to *topic 24*. *Topic 18* is clearly related to different types of neoplasms (abnormal mass in the tissue). Finally, *topic 20* and *topic 15* are related to kidney disease and liver disease respectively.

Summarization of Service/Location Topics: In Figure 4.6, We observe that *topic 00*, *topic 13* and *topic 19* are highly related to women who have been diagnosed with maternal conditions or delivery. The main reason is that they are at a very highly correlated location in the hospital i.e. prentice 12, prentice 8 Labor & Delivery. Also, they are either on obstetrics or gynecology service in the hospital. *Topic 03* and *topic 18* forms the thematic structure for orthopedics and neurology respectively. *Topic 02* and *topic 10* are highly biased towards locations treating patients with cardio problems.

Diagnosis	
topic 04	**prob**
delivery	0.19
cesarean	0.12
elderly multi	0.06
other maternal	0.06
sterilization	0.03
thyroid dysfunc	0.03
preech present	0.03
hypothyroidism	0.03
cord entanglement	0.02
group B strept.	0.02
topic 21	**prob**
delivery	0.17
abnormal heart rate	0.05
amniotic cavity	0.03
uterine inertia	0.03
hemorrhage	0.03
early delivery	0.03
fetal head	0.03
perineal laceration	0.02
cord entanglement	0.02
postterm pregnancy	0.02

Diagnosis	
topic 01	**prob**
pancytopenia	0.04
neutropenia	0.03
thrombocytopenia	0.03
stem replacement	0.03
diarrhea	0.02
fever	0.02
anemia	0.02
multiple myeloma	0.01
myeloid leukemia	0.01
candidiasis mouth	0.01
topic 17	**prob**
hypertension	0.12
hyperlipidemia	0.08
hypothyroidism	0.04
esophageal reflux	0.03
atrial fibrillation	0.03
type II diabetes	0.02
ischemic attack	0.02
osteoporosis	0.01
prostate	0.01
mental disorder	0.01

Diagnosis	
topic 18	**prob**
neoplasm of bone	0.04
neoplasm of liver	0.04
neoplasm of lung	0.03
pressure ulcer	0.03
constipation	0.03
neoplasm of brain	0.03
paraplegia	0.02
urinary infection	0.02
neoplasm of lung	0.01
palliative care	0.01
topic 24	**prob**
unspecified fall	0.03
occurrence, home	0.02
fall, slipping	0.02
collapse	0.02
other fall	0.01
hypertension	0.01
occurrence, other	0.01
unconscious	0.01
occurrence, street	0.01
open wound	0.01

Diagnosis	
topic 20	**prob**
kidney disease	0.11
chronic kidney S II	.11
kidney failure	.08
dehydration	0.05
type II diabetes	0.02
hyperlipidemia	0.02
hyperpotassemia	0.02
urinary infection	0.02
gout, unspecified	0.01
chronic kidney S III	0.01
topic 15	**prob**
cirrhosis of liver	0.06
other ascites	0.05
viral hepatitis C	0.03
alcoholic liver	0.02
liver transplant	0.02
diabetes mellitus	0.02
thrombocytopenia	0.02
kidney failure	0.01
encephalopathy	0.01
portal hypertension	0.01

Figure 4.3: Summarized Topics for Diagnosis

Procedure	
topic 02	**prob**
insert Endotracheal	0.14
venous Catheter	0.11
arterial Cath	0.10
cont Mech Vnt	0.09
cnt Mech Vnt 2	0.07
ext Infus Conc	0.03
temporary Trach	0.03
percutaneous Gast	0.03
endoscopy	0.02
trach Lavage	0.02
topic 03	**prob**
knee Replace	0.29
replace-Methacry	0.27
cell Tranfusion	0.10
anesth Injec	0.04
hip Bearing Surface	0.02
oth assisted	0.02
radiotherapeut	0.02
lap Appendectomy	0.01
transfusion Blood	0.01
aspiration Of Breast	0.01

Procedure	
topic 09	**prob**
fetus/Amnion	0.24
fetal Monitor	0.22
cervical c-section	0.21
induction Labor	0.13
artf. Membrane	0.10
fetal EKG	0.03
amnioinfusion	0.00
repair OB Uteri	0.00
influenza Vaccine	0.00
instr. Delived	0.00
topic 11	**prob**
manual Delivery	0.33
fetal Momitor	0.30
repair Laceration	0.25
artf. Membrane	0.11
breech Extraction	0.00
peritoneal Tiss	0.00
ob Vulva	0.00
skin Biopsy	0.00
rotate Fetal Head	0.00
alveolar Incision	0.00

Procedure	
topic 14	**prob**
cystoscopy	0.09
ureteral Cath	0.08
pyelogramy	0.05
breast Tissue	0.05
excise node	0.03
urine Incontion	0.03
vaginal Hyste	0.02
nephrostomy	0.02
mammoplasty	0.01
indwelling cath	0.01
topic 17	**prob**
extracorporeal	0.17
valv-Tissue	0.07
ultrasound-Heart	0.04
thor Ves Respect	0.03
valvuloplasty	0.03
cell Transfusion	0.03
venous Catheter	0.03
art Bypass	0.03
coronory Bypass	0.02
intercoastal Cath	0.02

Procedure	
topic 22	**prob**
cervical c section	0.39
fetal Monitor	0.36
artf. Membrane	0.07
bilat Tubal	0.05
therapeutic Aphere	0.02
vacuum Extract	0.01
tubal Destruct	0.00
indcut Labor	0.00
tamponade Uterus	0.00
Adhesiolysis	0.00
topic 23	**prob**
prostatectomy	0.10
cervical Node	0.07
robotic Procedure	0.06
lymph Node Exc	0.05
total Abdomin	0.04
nephroureterc	0.04
remove Tubes	0.04
laparoscopy	0.03
peritonial Tiss	0.03
remove Ovar	0.02

Figure 4.4: Summarized Topics for Procedures

4.5 Multidimensional Scaling

We employed a number of visualization techniques in order to see how the data was distributed in three dimensional space. With topic attributes of various dimensions, we reduced the dimensionality of the data to three di-

Medication	
topic 00	**prob**
clindamycin	0.32
gentamicin	0.18
ampicillin	0.11
aspirin	0.10
bupivacaine	0.06
docusate	0.06
ibuprofen	0.06
lactated injec	0.04
ondansetron	0.03
ferroussulfate	0.01
topic 03	**prob**
multivitamins	0.10
ascorbicacid	0.08
levetiracetam	0.07
docusate	0.06
aspirin	0.05
zincsulfate	0.04
esomeprazole	0.04
potassiumchloride	0.04
dexamethasone	0.04
oxacillin	0.04

Medication	
topic 04	**prob**
vancomycin	0.08
multivitamins	0.08
warfarin	0.07
docusate	0.07
cefazolin	0.07
bupivacaine	0.06
esomeprazole	0.06
aspirin	0.05
ferrousgluconate	0.05
celecoxib	0.04
topic 10	**prob**
ketorolac	0.14
cefazolin	0.12
lactated ringers inj	0.11
aspirin	0.10
ibuprofen	0.10
docusate	0.10
simethicone	0.08
tripedia	0.05
bupivacaine	0.04
ondansetron	0.02

Medication	
topic 17	**prob**
trimethoprim	0.08
tacrolimus	0.08
mycophenolatem	0.06
prednisone	0.06
methylprednis	0.04
famotidine	0.03
valganciclovir	0.03
aspirin	0.03
insulin	0.03
bupivacaine	0.02
topic 14	**prob**
magnesiumsulfate	0.05
morphine	0.05
furosemide	0.05
metoprolol	0.04
glucose	0.04
heparin	0.04
insulin	0.04
potassiumchl	0.03
vancomycin	0.03
docusate	0.03

Medication	
topic 18	**prob**
polyethylene	0.11
docusate	0.10
bisacodyl	0.08
simethicone	0.06
bupivacaine	0.05
diazepam	0.04
aspirin	0.04
magnesiumcitrate	0.03
vancomycin	0.03
lactulose	0.02
topic 09	**prob**
ibuprofen	0.15
oxytocin	0.15
docusate	0.14
lactated inject	0.14
tripedia	0.10
aspirin	0.10
bupivacaine	0.05
penicillin	0.05
influenza vac.	0.02
rhodimmuneglob	0.02

Figure 4.5: Summarized Topics for Medications

mensions, such that distances between points in the higher dimensional topic space were preserved. The easiest method for doing this was to find the 3 principle components of the data, via principle component analysis (PCA) [22], and plot points along these vectors. While easy to implement, this only de-correlated dimensions. Another, more advanced approach we considered was Multi-Dimensional Scaling (MDS) [23] [24]. MDS consists of a suite of dimensionality reduction and scaling techniques for preserving inter point distances as much as possible; it comes in two general forms, metric and non-metric. An MDS algorithm starts with a matrix of item - item similarities, then assigns a location to each item in N-dimensional space, where N is specified a priori. [23] For sufficiently small N, the resulting locations may be displayed in a graph or 2D visualization techniques such as scatter plots. However, not all dataset were capable of being transformed using this method; in such instances, we used PCA to transform the data. It should be noted that the data operated on was already transformed at the time of visualization via LDA; thus, dimensions were already fairly independent.

MDS has now become a general data analysis technique used in a wide variety of fields [24]. MDS pictures the structure of a set of objects from data that approximates the distances between pairs of the objects. Each object or event is represented by a point in a multidimensional space. The points are arranged in this space so that the distances between pairs of points have the strongest possible relation to the similarities among the pairs of objects.

Service/Location		Service/Location		Service/Location		Service/Location	
topic 00	**prob**	**topic 03**	**prob**	**topic 10**	**prob**	**topic 18**	**prob**
prentice 12	0.30	orthopedics	0.33	fienberg 11w	0.3	fienberg 10w	0.25
prentice 8 l&d	0.28	fienberg 14w	0.27	cardio thoracicsurg	0.18	neurosurgery	0.19
obstetrics	0.27	asu recovery 65	0.27	fienberg 7e cticu	0.18	nicu	0.15
prentice triage	0.12	fienberg 10 se	0.08	asu recovery 65	0.18	neurology	0.12
nst triage	0.00	padm	0.01	vascular surgery	0.18	fienberg 10se	0.10
fienberg 8e sicu	0.00	fienberg 8e sicu	0.00	fienberg 8e sicu	0.09	asu recovery 65	0.09
communicable	0.00	plastic surgery	0.00	padm	0.04	emergency dept 1	0.04
pediatrics	0.00	oa surg	0.00	fienberg 15e	0.01	padm	0.01
gynecology	0.00	nicu	0.00	ccu	0.00	fienberg 10 necrc	0.00
outp	0.00	fienberg 13 w	0.00	fienberg 10se	0.00	et	0.00
topic 02	**prob**	**topic 08**	**prob**	**topic 13**	**prob**	**topic 19**	**prob**
general medicine	0.18	fienberg 12w	0.21	prentice 14	0.36	gynecology	0.36
fienberg 15e	0.18	general surgery	0.19	prentice asu	0.25	prentice 8 l&d	0.29
cardiology	0.18	emergency dept 1	0.16	gynecology	0.15	prentice triage	0.13
fienberg 14e	0.16	fienberg 12e	0.07	gyneoncology	0.11	prentice 13	0.10
ccu	0.10	fienberg 8e sicu	0.07	plastic surgery	0.05	prentice 11	0.10
emergency dept 1	0.07	surgery endocrine	0.07	surgical oncology	0.02	nst triage	0.01
fienberg 15w	0.03	edm	0.06	fienberg 8e sicu	0.00	ccu	0.00
micu	0.03	fienberg 11e	0.02	emergency dept 1	0.00	pediatrics	0.00
nicu	0.03	nicu	0.02	radiology	0.00	icr	0.00
csu	0.00	surgical oncology	0.01	edm	0.00	prentice 12	0.00

Figure 4.6: Summarized Topics for Service/Locations

That is, two similar objects are represented by two points that are close together, and two dissimilar objects are represented by two points that are far apart. The space is usually a two- or three-dimensional Euclidean space, but may be non-Euclidean and may have more dimensions [25]. We have implemented two types of MDS methods i) Metric Classical MDS ii) Non-Metric Classical MDS [25] defined as follows:

i) Metric multidimensional scaling : A superset of classical MDS that generalizes the optimization procedure to a variety of loss functions and input matrices of known distances with weights and so on. A useful loss function in this context is called stress, which is often minimized using a procedure called stress majorization.

ii) Non-metric multidimensional scaling : In contrast to metric MDS, non-metric MDS finds both a non-parametric monotonic relationship between the dissimilarities in the item-item matrix and the Euclidean distances between items, and the location of each item in the low-dimensional space. The relationship is typically found using isotonic regression.

Figure 4.7 illustrates the results of MDS on the user typing in section 4.3. The first Figure shows the distribution of all the users in 3D space based on their access to particular diagnosis topic (derived from LDA). The green

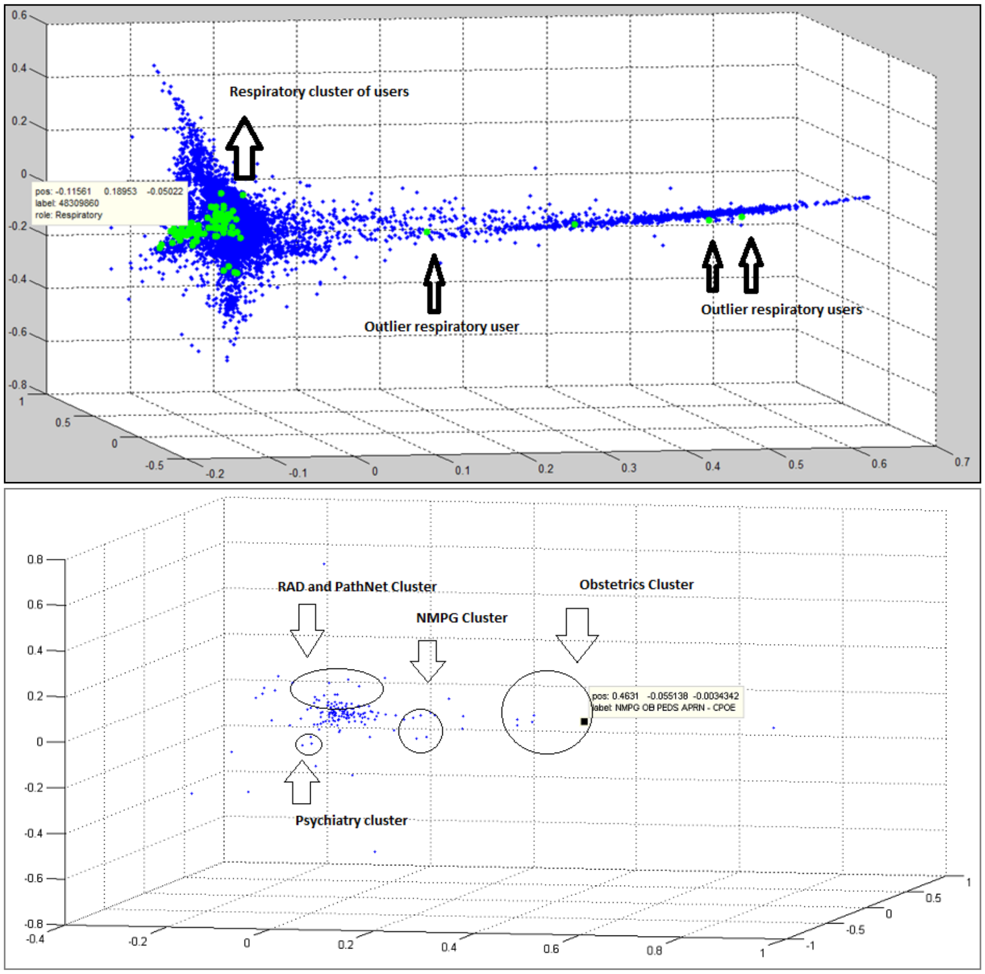


Figure 4.7: User Typing Visualization for Diagnosis

dots are the users within same role, whereas the blue dots are all the users irrespective of the role. For instance, the users who access patients having diagnosis neurological disorders will be at less distance than another user accessing patients of type obstetrics. In 4.7 we can see that the green dots represent clusters of users having role respiratory, but some users of the same role are at a far distance from the original cluster, which may be an outlier. In Chapter 5, we develop a framework to detect such outliers in the system. The figure also shows the roles cluster together whose users behave in a similar way. Hence, we can find RAD cluster, Psychiatry cluster, Obstetrics cluster etc. which also give us an insight into combining different similar roles dynamically based on the access of the users.

Chapter 5

Anomaly Detection Model

Anomaly Detection is the process of finding outliers from a given data set. According to the anomaly detection survey [26] the techniques can be grouped into following categories: classification based, nearest-neighbor based, clustering based and statistics based. Classification based algorithms are mainly supervised algorithms that assume distinction between anomalous and normal instances can be modeled for a particular feature space. For instance, given a user and label (normal or anomalous), we can learn a binary classifier to predict if the new user is anomalous or not based on the feature space. Whereas, nearest-neighbor based algorithms assume that anomalies lie in sparse neighborhoods and that they are distant from their nearest neighbors. They are mainly unsupervised algorithms. One example of such algorithm is k-nearest neighbor algorithm which is explained in section 5.1. Clustering based algorithms work by grouping similar objects into clusters and assume that anomalies either do not belong to any cluster, or that they are distant from their cluster centers or they belong to small and sparse clusters. At last, the statistical techniques such as chi-square analysis can be done to compare the expected distribution over observed distribution and if they differ significantly then we can detect outliers. The intuition is that if the confidence level of any attribute in the observed distribution is significantly lower than in the expected distribution, then it is termed as anomalous. The chapter is organized as follows: section 5.1 discusses the mathematical model behind the k-nearest neighbor technique and how we can use it in our RTAD framework. In section 5.2 we formalize our model by defining various anomalous users for detection based on k-nearest neighbor and LDA, discussed in section 4.1.

5.1 *k*-Nearest Neighbor

An anomalous user is a user whose distance from its nearest neighbors is sufficiently large with respect to the average distance between the neighbors. Distance in our definition is defined as the Euclidean distance between different user points in space. Users with the same role are plotted according to their conditional attribute values or topic distributions: $P(D_P \mid U)$. The topic probability distribution has been derived in section 4.3 where we derived the user topics from the patients based on the accesses of the user. For the purposes of determining anomalous users, we use the k nearest neighbor's algorithm, which tags outliers with the same definition we gave to anomalous users. To formulate this, suppose the set of nearest neighbors for user U is N(U). Let the average distance between nearest neighbors be defined as the following:[1]

$$d_{NN} = \sum_{n,m \in N(U), n \neq m} \frac{dist(n,m)}{\frac{K(K-1)}{2}} = \frac{2}{K^2 - K} \sum_{n,m \in N(U), n \neq m} dist(n,m) \tag{5.1}$$

Similarly, let the average distance between the user and each neighbor be defined as the following:

$$d_{NU} = \frac{1}{K} \sum_{i=1}^{K} dist(u, n_i) \tag{5.2}$$

Given this, a user is recorded as anomalous according to the following piecewise equation, where β is the chosen threshold.

$$f(U,K) = \begin{cases} anomalous & \frac{d_{NN}}{d_{NU}} > \beta \\ regular & otherwise \end{cases}$$

5.2 Random Topic Access Model

The Random Topic Access (RTA) model is a framework for describing anomalous users in terms of random topics, as opposed to random access patterns.

[1] There are $\frac{K(K-1)}{2}$ pairs of K nearest neighbors

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 this framework, RTAD. Considering the Dirichlet distribution is the conjugate prior of the multinomial, we felt it appropriate to convey our argument with respect to the concentration parameter, α. Figures 5.2, 5.3, 5.4 use MDS to project all the users in the system based on their diagnosis information (derived in user typing). We then sample 5% of the directed users from the Dirichlet distribution and insert into the system to visualize the overall distribution of normal users with respect to directed users.

5.2.1 Dirichlet Distribution

Let us first get an insight into the Dirichlet distribution and its concentration parameter α. The Dirichlet distribution [27] is a family of continuous multivariate probability distributions parametrized by a vector α of positive reals. It is the multivariate generalization of the beta distribution. Dirichlet distributions are very often used as prior distributions in Bayesian statistics, and in fact the Dirichlet distribution is the conjugate prior of the categorical distribution and multinomial distribution. The space of all m-dimensional multinomials is an $(m-1)$-simplex by definition, and so the Dirichlet distribution can also be thought of as a distribution over a simplex.

Algebraically, the distribution is given by

$$Dir(\mathbf{p}|\alpha_1, \ldots, \alpha_m) = \frac{1}{Z} \prod_k p_k^{\alpha_k - 1}$$

where $Z = \frac{\prod_{k=1}^{m} \Gamma(\alpha_k)}{\Gamma\left(\sum_{k=1}^{m} \alpha_k\right)}$ is a normalization factor. [2]

There are m parameters α_k which are assumed to be positive. Figure 5.1 in [28] shows the density plots (blue=low, red=high) for the Dirichlet distribution over the probability simplex in 3D for various values of the parameter α. The author in [29] further explains that when α = [k, k, k] for some k > 0, the density is symmetric about the uniform pmf (which occurs in the middle

[2] $\Gamma(x)$ denotes the Gamma function and is defined to be: $\int_0^\infty t^{x-1} e^{-t} dt$.

of the simplex), and the special case α = [1, 1, 1] shown in Figure 5.1 part 3 and has the uniform distribution over the simplex. In our RTAD model, this is analogous to selecting topics (say 3 topics) from the simplex such that the probability of selecting each topic is completely random and unbiased towards any part of the simplex. When 0 < k < 1, there are sharp peaks of density almost at the vertices of the simplex and the density is concentrated at the corners of the simplex, as seen in 5.1 part 2. Hence, the probability of selecting each topic in this case is not completely random and is biased towards the corners of the simplex having high density points. When k > 1, the density becomes concentrated in the center of the simplex, as shown in the Figure 5.1 part 1.

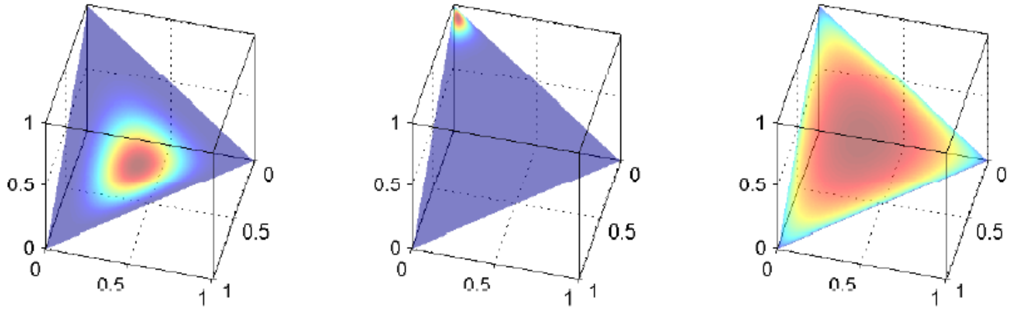


Figure 5.1: Density Plots for Dirichlet Distribution $\alpha > 1$, $\alpha < 1$, $\alpha = 1$

5.2.2 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. Differentiating the anomalous user from the real user are the topics ascribed to the anomalous user's access patterns. While these topics may be ordinary with respect to the hospital population, it could be deviant with respect to the population of users similar to the real user. The anomalous user in this case could be sampled from a Dirichlet with $\alpha < 1$, since real users are assumed to be strongly biased towards a set of few topics. Given a real user with an typical type of topic distribution, it is highly probable that generating random anomalous users will result in anomalous users not biased towards the same topic as the real user.

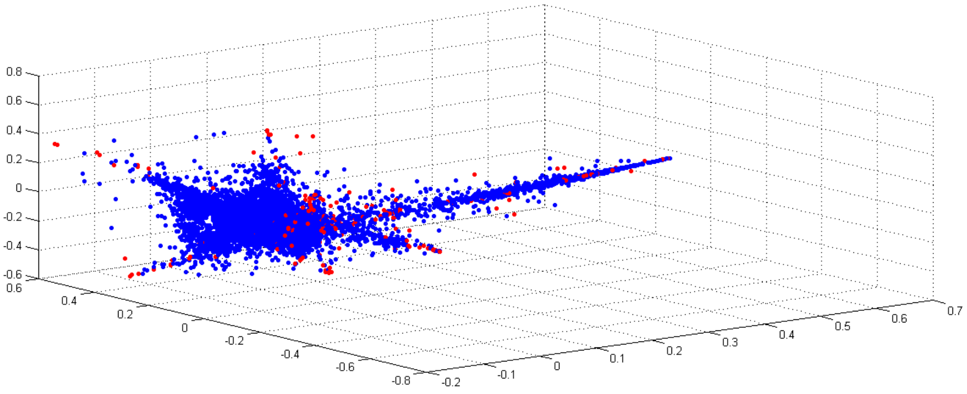


Figure 5.2: Injected Directed Users for $\alpha < 1$

5.2.3 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 RAO 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 nice property of allowing the system to be tested against input that has not even been seen in the data yet.

5.2.4 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 for everyone to use; out of sheer

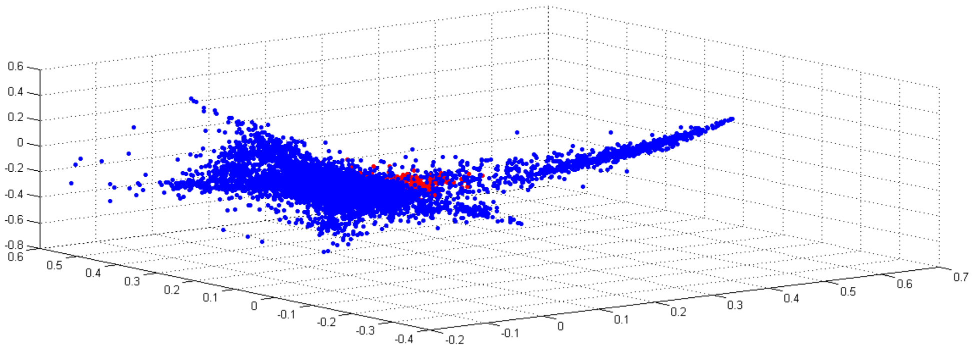


Figure 5.3: Injected Completely Random Users for $\alpha = 1$

convenience, users of many different specializations log in and make accesses under this user's account. Because the resulting accesses are made by many different kinds of users, the logged in user resembles a sort of average of these different extreme values. This 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.

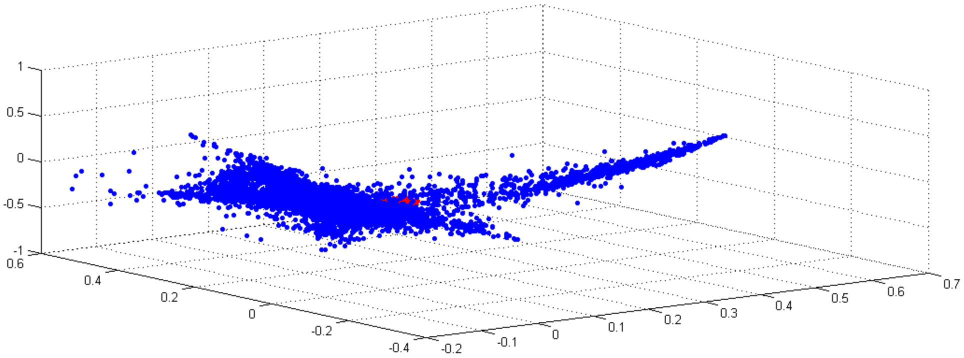


Figure 5.4: Injected Indirect Users for $\alpha > 1$

With these different types of random users modeled in our system, we move to our implementation of the RTA model.

Chapter 6

Experimental Results and Evaluation

Our RTA framework, RTAD, consists of running LDA on the entire population of users, typing users with respect to their accesses and patients, identifying the 5 most populated user roles, and 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 whole user population. A visualization of the effect of these different α values on one of these roles is given in Figure 6.1. Utilizing classical multidimensional scaling for dimensionality reduction, we graphed the projections of the high dimensional topic space for NMH Resident Fellow CPOE and different α such that the pairwise distance between users was preserved. Real users are shown in blue while anomalous users are shown in red. As can be seen comparing the top plots to the bottom plots, α values less than 1 result in a greater amount of dispersion with respect to real users than $\alpha >= 1$. As α increases, the random users become more and more clustered, making the anomaly detection more difficult. This chapter is organized as follows: section 6.1 describes the basic statistics about the roles in the experiment and the distribution of users within the roles. Sections 6.2, 6.3 and 6.4 analyze anomaly detection experiments for all five roles, for different values of α. Section 6.5 summarizes the analysis and discusses the intuition behind the results.

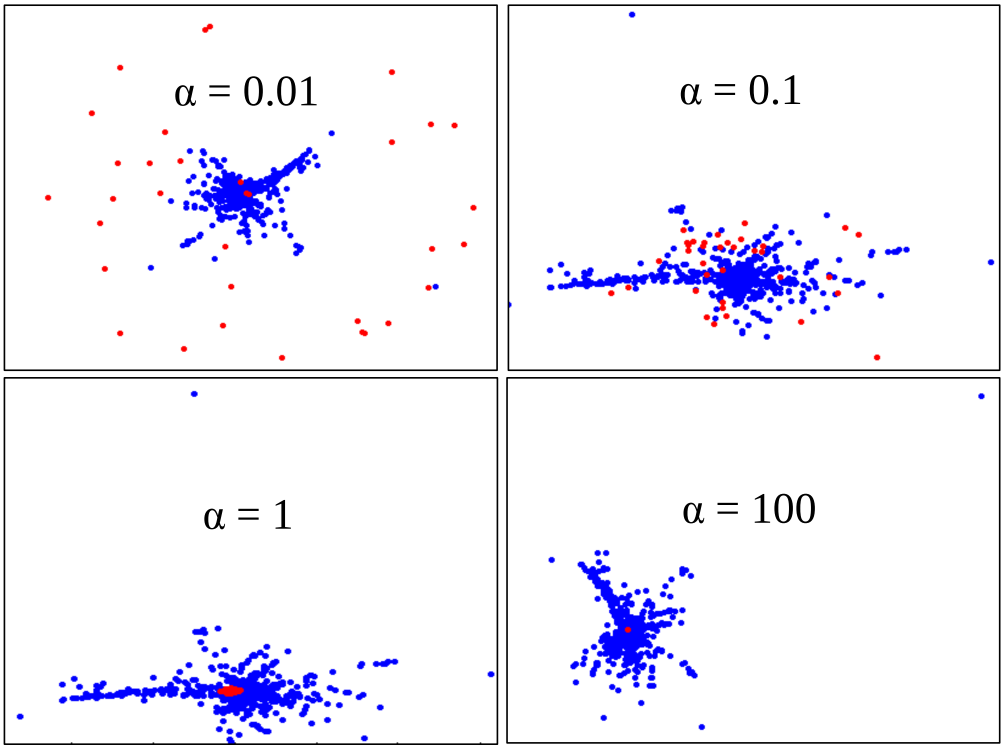


Figure 6.1: NMH Resident Fellow CPOE (mixed topics)

6.1 Role Description

In Chapter 3 we observed the sparse nature of the dataset where very few roles have very large number of accesses and most of the roles have few accesses and users with the role. For the same reason, we decided to do the analysis on five most populated roles with respect to number of users within that role in the NMH dataset. Note that it does not change any aspect of the analysis as the experiments are independent of the role chosen and can be done on any given role. Basic description of the roles selected in our analysis is as follows:

1. Med Student CPOE : All the users in this role made 149,683 accesses in total of four month time period and the total number of users within this role is 475. The mix rate of 5% random users gives 23 random users.

2. NMH Resident Fellow CPOE : All the users in this role made 722,137 accesses in total of four month time period and the total number of

users within this role is 710. The mix rate of 5% random users gives 35 random users.

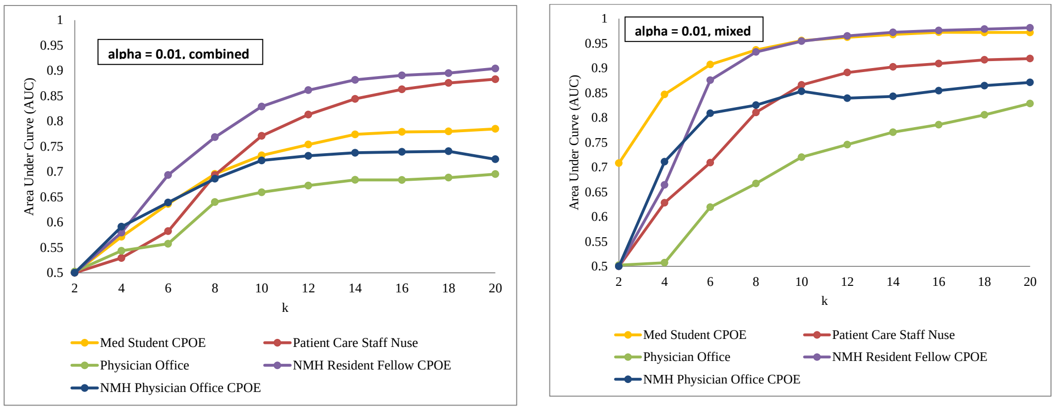
3. NMH Physician Office CPOE : All the users in this role made 48,305 accesses in total of four month time period and the total number of users within this role is 250. The mix rate of 5% random users gives 12 random users.

4. Physician Office : All the users in this role made 68,111 accesses in total of four month time period and the total number of users within this role is 422. The mix rate of 5% random users gives 22 random users.

5. Patient Care Staff Nurse : All the users in this role made 1,841,851 accesses in total of four month time period and the total number of users within this role is 1429. The mix rate of 5% random users gives 71 random users.

6.2 Simulating Directed or Masquerading User

RTAD aims at detecting three type of anomalous users and the first kind is defined as a directed user, discussed in 5.2. In this scenario, an anomalous user of a particular speciality 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. In Figure 6.3 we plot the AUC curves to detect the number of directed users captured by RTAD. In 6.3 all the users are taken into consideration without any role information. We repeat the k-NN experiments on the users having topic distributions from diagnosis, medications, service/location, procedures, mixed and combined. We observe that the model performs well for the values of $k > 8$ in the case of service/location and for others it increases exponentially. The best dimension observed is the mixed dimension to capture directed users in the system which is ignorant of role information. As we see that the combined and mixed dimension performs equally well in the case of all users, we picked these two dimensions to do the k-NN experiments for each of the five roles shown in Figure 6.2(a) and in Figure 6.2(b). The



(a) Combined Dimension

(b) Mixed Dimension

Figure 6.2: Directed User (α=0.01)

analysis shows that for the same dimension and the values of k the AUC is different in each role. We observe that for $k > 8$ the AUC value for both mixed and combined dimension, in the case of NMH Resident Fellow CPOE is the best among all the roles. One reason for this might be that the role NMH Resident Fellow CPOE is tightly clustered with respect to other users, making the anomaly detection easier. In summary, to detect directed users, RTAD performs the best in case of mixed dimension for all users and NMH Resident Fellow CPOE, given the role information about the users. On comparison, mixed dimension is the best as compared to combined dimension for directed RTAD model.

6.3 Simulating Pure Random User

Another type of user RTA model can handle is a pure random user. We have already discussed this type of user in 5.2. This type of user is characterized by completely random behavior, with little semantic congruence to the hospital setting. By generating random users from a Dirichlet with $\alpha = 1$, any type of random user can be generated. In Figure 6.5(a) we perform the k-NN experiments with respect to all the users in the system and in the absence of role information. We repeat the experiment for all given dimensions of the patient with respect to the user. In this case, we observe that the mixed dimension performs the best as compared to other dimensions to detect a

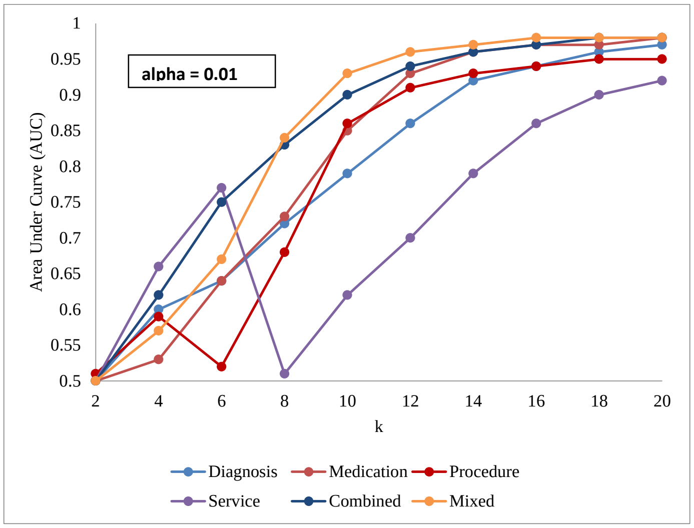
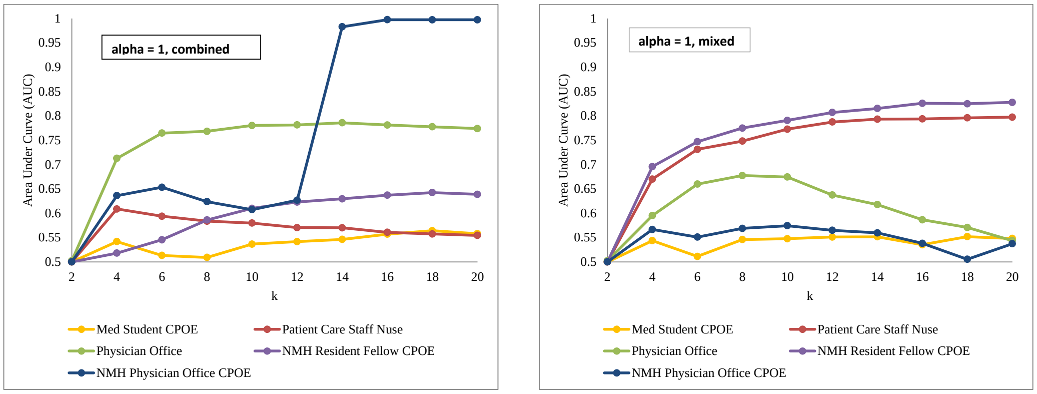


Figure 6.3: AUC Across all Users for $\alpha = 0.01$

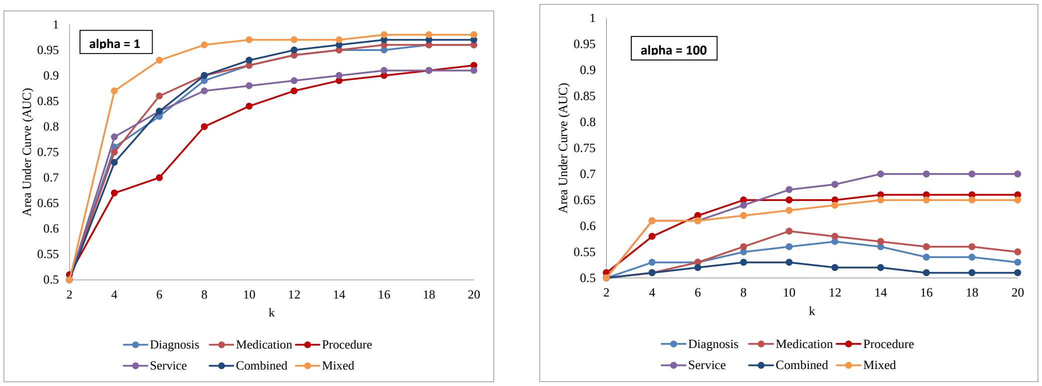
purely random user. The reason seems to be the tightly clustered users in the case of mixed topics and hence, easier to detect anomalous users. Now, we perform the same analysis taking combined and mixed dimensions and also informing the system about the role of the users. In Figures 6.4(a) and 6.4(b), we observe that in case of mixed dimension, Med-student CPOE performs the worst and NMH Resident Fellow CPOE performs the best. One reason for bad performance of a medical student might be the non-clustering of the medical students based on their access to the patients. As medical students are responsible for accessing different themes in the hospital, they seem like random user which makes them difficult to differentiate from the anomalous user. On the other hand, in the case of NMH Resident Fellow CPOE it is the opposite. But, in the case of combined dimensions, NMH Physician Office CPOE performs the best for values of $k > 12$, anomalous users appearing more clustered with respect to real users. In summary, mixed dimension performs the best when the system has no information about the roles, while combined seem to perform better in the case of NMH Physician Office CPOE, but otherwise it is consistently good in the case of mixed dimension.



(a) Combined Dimension

(b) Mixed Dimension

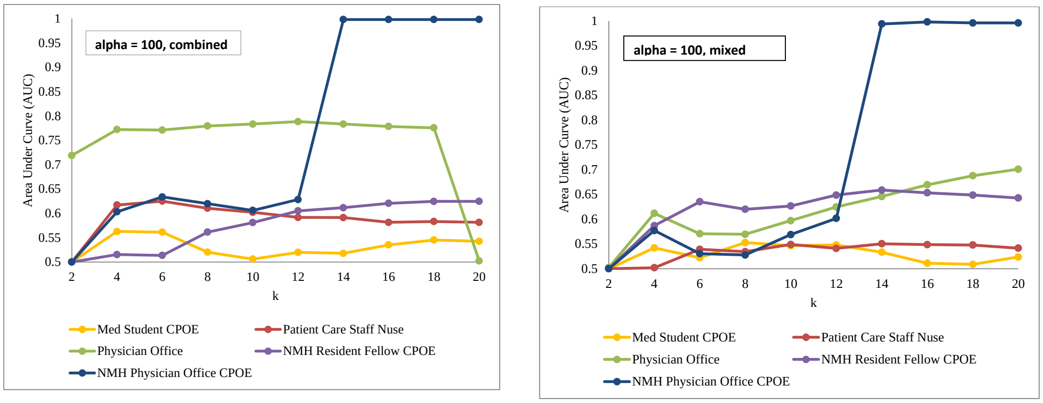
Figure 6.4: Pure Random User (α=1)



(a) α = 1

(b) α = 100

Figure 6.5: AUC across all users for α=1 and α=100



(a) Combined Dimension

(b) Mixed Dimension

Figure 6.6: Indirect User (α=100)

6.4 Simulating Indirect User

The third type of user modeled by RTA is the indirect user. The definition of this user is also given in 5.2. 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 access to his terminal open for everyone to use; out of sheer convenience, users of many different specializations log in and make accesses under this users account. This anomalous user can best be modeled with $\alpha > 1$ in the Dirichlet distribution. In Figure 6.5(b), we perform the k-NN experiments with respect to all the users in the system and in the absence of role information. We repeat the experiment for all given dimensions of the patient with respect to the user. We observe that the system performs very poor in detecting indirect user. The reason is that the random users injected in the system are clustered together and hard to detect as an outlier. Service/Location performs as the best dimension among others to best differentiate indirect RTA user from the other users. In Figures 6.6(a) and 6.6(b), we observe that in the case of mixed dimension, almost all the roles perform bad except NMH Physician Office CPOE and the performance increases for the values of $k > 12$. NMH Physician Office CPOE shows similar results in the case of combined dimension. We will explore the reason for this inversion in the next section. In general medical student CPOE seems to be performing the worst among all the roles. Though in the case of indirect users, both combined and mixed dimension perform equally bad.

6.5 Summary of the Analysis

In summary, we compared the best AUC for each role and α. Figure 6.7 and Figure 6.8 show the best AUC's for each role-α combination, with the whole user population used as a control. For masquerading users $\alpha < 1$, the resulting AUC's are extremely strong, especially for highly specialized users ($\alpha = 0.01$). This is expected; since 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 result, they will approach the maximum distance that can be achieved on the simplex and will appear

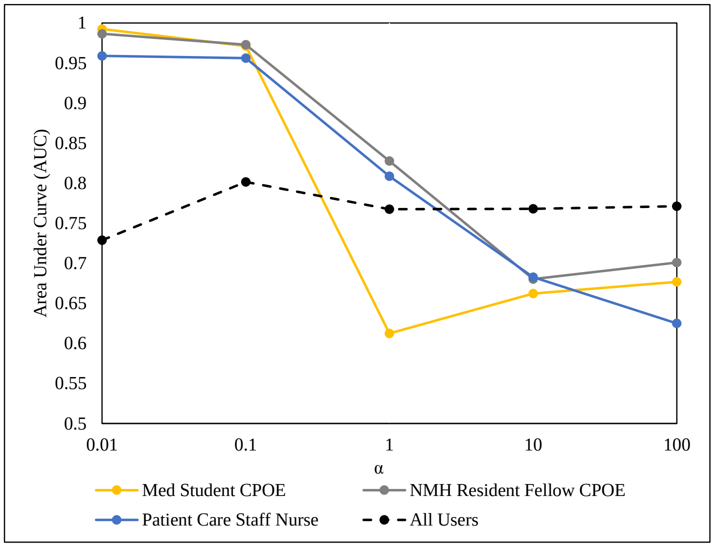


Figure 6.7: The best AUC across all evaluated dimensions is plotted for each role performing badly for $\alpha > 1$.

more varied with respect to the users in the role. As the system transitions to more random users, the resulting AUC's suffer somewhat for all roles, except for NMH Physician Office CPOE. In analyzing the results of NMH Physician Office CPOE, the system actually inverts itself, with anomalous users appearing more clustered with respect to real users. This trend continues as the system is evaluated against undirected users, which is also expected, as the anomalous users will become more and more clustered. With respect to the baseline, utilizing semantic role information is a huge boost to the system for directed users and generally performs as well or better for purely random users. Performance suffers for some roles tested against undirected users compared to the baseline; this discrepancy is intuitive in the context of k-NN as the simplex is more populated in the baseline case, meaning that there is a higher likelihood of local clusters of users across different roles. With respect to the Med-Student CPOE role, our findings regarding the response of this role to the RTA framework 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, they will have accummulated many different kinds of accesses into their history. As a result, it would be expected that this role would suffer the most when tested against the purely random user, as this is what the average medical student could be modeled as. Additionally, it is no surprise that the results

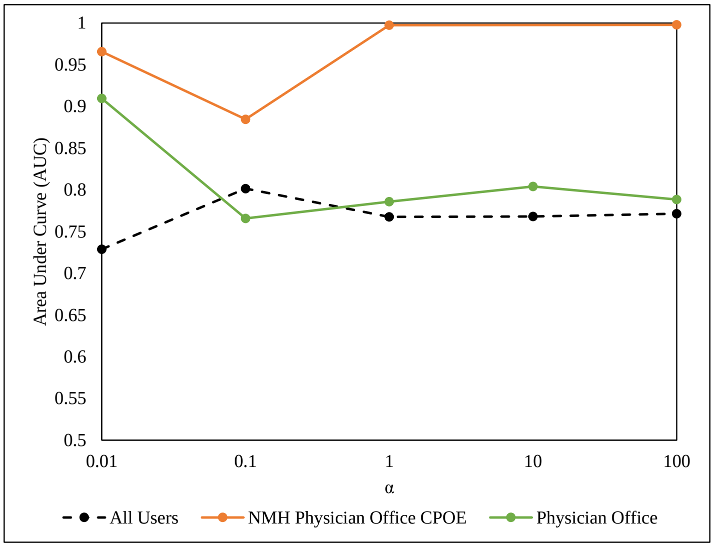


Figure 6.8: The best AUC across all evaluated dimensions is plotted for each role performing well or near average for $\alpha > 1$.

recover somewhat as the anomalous users become more tightly clustered, but less random.

Regarding our inquiry of which dimension outperforms others with respect to AUC, there is not any clear advantage from utilizing mixed information versus concatenated vectors or mixed topics versus single dimension topics. The best topic dimensions selected for each role - α varied considerably. So much so, that no discernible trend could be reached from this small dataset.

Chapter 7

Conclusions and Future Scope

Overall, we were able to demonstrate a lack of coverage in the existing methodology for evaluating security models utilizing random users. The classical technique, modeling atypical access by completely random behavior (ROA), is constrained by the particular dataset and can not necessarily imagine all types of conceivable attackers. Utilizing latent topics models such as LDA, the RTA model provides more robust coverage of the different type of attackers by generating synthetic users directly from a topic simplex, as opposed to data. In this manner, you can think of the dataset as being a sample from a larger, unseen population distribution. Transformation to the topic domain may not allow us to realize 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, controlling 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. We also plan to extend the RTAD framework by using Labeled LDA (supervised learning algorithm), which learns the model by the information given in the form of patient labels. The intuition is that the cluster of users formed in this case will be more compact than in unsupervised LDA and hence it will be easier to detect anomalies in such a framework.

References

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

[2] [Online]. Available: http://www.hhs.gov/ocr/privacy/

[3] S. Gupta, C. Hanson, C. A. Gunter, F. Mario, D. Liebovitz, and B. Malin, "Modeling and detecting anomalous topic access," *IEEE International Conference on Intelligence and Security Informatics, 2013*, 2013.

[4] D. Garets and M. Davis, "Electronic medical records vs. electronic health records: Yes, there is a difference," *HIMSS Analytics White Paper*, pp. 772–776, 2006.

[5] W. R. Herrsh, "The electronic medical record: Promises and problems," *Journal of the American Society for Information Science*, 1995.

[6] [Online]. Available: hhttp://www.whitehouse.gov/sites/default/files/microsites/ostp/pcast-health-it-report.pdf

[7] G. Carl, D. M. Liebovitz, and B. Malin, "Experienced based access management," *IEEE Computer and Reliability Societies*, vol. 3, pp. 48–55, 2011.

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

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

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

[11] 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.

[12] 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.

[13] Y. Chen, S. Nyemba, W. Zhang, and B. Malin, "Leveraging social networks to detect anomalous insider actions in collaborative environments," *IEEE International Conference on Intelligence and Security Informatics*, pp. 119–124, 2011.

[14] 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.

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

[16] 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.

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

[18] R. Chinchani, A. Muthukrishnan, and S. Upadhyaya, "Racoon: Rapidly generating user command data for anomaly detection from customizable templates," *20th Annual Computer Security Applications Conference*, 2004.

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

[20] W. Van der Aalst, T. Weijters, and L. Maruster, "Workflow mining: discovering process models from event logs," *IEEE Transactions on Knowledge and Data Engineering*, 2004.

[21] B. van Dongen, A. de Medeiros, H. Verbeek, A. Weijters, , and W. van der Aalst, "The prom framework: a new era in process mining tool support," *ICATPN'05 Proceedings of the 26th international conference on Applications and Theory of Petri Nets*, pp. 444–454, 2005.

[22] [Online]. Available: https://en.wikipedia.org/wiki/Principal_component_analysis

[23] [Online]. Available: http://en.wikipedia.org/wiki/Multidimensional_scaling

[24] M. L. Schiffman. S. S Reynolds and Y. F. W., "The prom framework: a new era in process mining tool support," *Introduction to Multidimensional Scaling. Academic Press, New York*, 1981.

[25] [Online]. Available: http://forrest.psych.unc.edu/teaching/p208a/mds/mds.html

[26] V. Chandola, A. Banerjee, , and V. Kumar, "Anomaly detection: A survey. technical report," 2007.

[27] [Online]. Available: http://en.wikipedia.org/wiki/Dirichlet_distribution

[28] J. Huang, "Maximum likelihood estimation of dirichlet distribution parameters."

[29] B. A. Frigyik, A. Kapila, and M. R. Gupta, "Introduction to the dirichlet distribution and related processes," *UWEE Technical Report Number UWEETR-2010-0006*, 2006.