A FRAMEWORK FOR THE AUTOMATIC DISCOVERY OF POLICY FROM HEALTHCARE ACCESS LOGS

By

John M. Paulett

Thesis

Submitted to the Faculty of the

Graduate School of Vanderbilt University

in partial fulfillment of the requirements

for the degree of

MASTER OF SCIENCE

in

Biomedical Informatics

August, 2009

Nashville, Tennessee

Approved:

Professor Bradley Malin

Professor Dario Giuse

Professor Nancy Lorenzi

ACKNOWLEDGEMENTS

I would like to deeply thank my advisor, Dr. Brad Malin, and my committee, Dr. Nancy Lorenzi and Dr. Dario Giuse, for their guidance and support of my master's research. This research was funded in the past by NSF grant CCF- 0424422, the Team for Research in Ubiquitous Secure Technologies. I appreciate the assistance of Dr. Edward Shultz and the Vanderbilt Informatics Center in obtaining my degree.

This work has benefited from discussions with numerous members of Vanderbilt University, including Dr. Randolph Miller, Dr. Cynthia Gadd, Dr. Josh Peterson, Dr. Dominik Aronsky, Gaye Smith, Dr. Joshua Denny, Karen Hughart, and Dr. Nancy Wells.

The access logs used in this study were provide by Dr. Dario Giuse and David Staggs. I also received data from Vanderbilt's Enterprise Data Warehouse.

This work was made possible by a number of open-source or free software tools, including: the Python language, the R language, ggplot2, numpy, matplotlib, NetworkX, Gliffy, and Jude.

I thank the anonymous physicians who respondent to my survey as well as the RedCap survey system, funded by NCRR/NIH grant 1 UL1 RR024975.

Finally, I thank Laura Duvall for her support and assistance editing this manuscript.

TABLE OF CONTENTS

Page

LIST OF TABLES

Page

LIST OF FIGURES

Page

CHAPTER I

INTRODUCTION

Healthcare Organizations (HCOs) are inherently complex bodies of clinicians and support staff working to provide care for patients (1; 2). In addition to this inherent complexity, HCOs deal with constant change, whether adopting new protocols, reacting to legislated changes, providing care in emergency situations, or annually introducing new classes of medical students and residents. The complexity and constant change of the organizational structure and workflows makes HCOs difficult to accurately model using traditional techniques. This difficulty particularly stymies the efforts of HCO privacy officials who wish to increase system security and ensure patient confidentially by applying fine-grained access control systems to the electronic Health Information Systems (HIS) within the HCO, systems such as Electronic Medical Record Systems (EMRS).

When security experts discuss healthcare, they often suggest that healthcare systems use the security models introduced by the military and by the banking industry. These groups have effectively implemented security measures such as role-based access control (3; 4), in which users are assigned a role or set of roles, which determines his or her access to information.

Unfortunately, these access control systems are not widely adopted by HCOs for several reasons. First, clinicians have an understandable fear of being encumbered by

inappropriately configured access rights, making it difficult to access critical patient information in a timely fashion, particularly in medical emergencies (3). Second, the organizational complexity of HCOs is incredibly difficult to capture accurately or completely using the traditional top-down approach of manually defining the access policy. In a typical top-down approach, security experts try to manually codify the allowed actions within the system, as well as assign these actions to the proper individuals. Unfortunately, this type of modeling can suffer from issues of informant accuracy (5). Third, the dynamic nature of an HCO can quickly invalidate manually-defined access policies.

Since the policy definitions of access control systems can be too burdensome for HCOs, many fall back to retrospective auditing of users' actions within the system (6). Unfortunately, auditing merely shifts the manual burden from upfront policy definitions to the massive task of combing through thousands or millions of user actions every day. In practice, auditing reduces to random spot checks and requested reviews. Even at advanced institutions, consequently, the level of auditing sophistication is very basic.

Currently, many clinical systems are essentially vulnerable to insider attacks (attacks in which the user has a legitimate right to access the system, but not necessarily rights to specific information within the system) (7).

The lack of adoption of access control systems or thorough auditing within HIS is due neither to inherent problems in these technologies nor to the lack of vendor implementations within HIS products. Rather, this lack is caused by the inability to create appropriate policy definitions of what is normal and abnormal usage of the HIS.

In this work, we will demonstrate a novel tool that can model an HCO using a bottom-up approach. Our approach constructs a statistical model of the HCO by analyzing the usage of a HIS through its access logs.

To our knowledge, no work exists that attempts to mine policies from access logs in the HCO setting. Previous work, however, has demonstrated the viability of using access logs to characterize user behavior (8; 9). Additionally, the fields on which the methods are based, such as social networking and data mining, have a successful history of distilling data and providing meaningful results.

Our methods and tool are designed so an HCO can generate a statistical model of its users' interactions. This statistical model can form the basis of policy definitions and rules in a real-world access control system or auditing system. Specifically, our work will feed into the efforts in model-based software platforms, such as Vanderbilt's Model-Integrated Clinical Information System, a system that assists in the rapid development and evaluation of formal systems based on service oriented architectures (10; 11). These platforms have integrated robust privacy and security policy specification and validation languages, such as Stanford's logic based on contextual integrity (12). The statistical modeling technique presented in this paper provides the dynamic policies needed by these systems.

In this thesis, we first cover the issues with existing solutions and explain how this work relates to other work in the field. Second, we present an open-source framework, the Healthcare Organization Relational Network Extraction Toolkit (HORNET), that allows HCO officials to generate a statistical model based upon their

own information systems and allows researchers to build on top of our work. Third, we explain an approach that HORNET can use to generate a statistical model of how the organization works from the organization's access logs. Using this approach, we next demonstrate HORNET by generating a statistical model from the access logs of Vanderbilt University Medical Center's (VUMC) EMRS, known as StarPanel. With a statistical model built from StarPanel's access logs, we conduct a pilot survey to demonstrate the experts' inability to model how the organization works. We conclude this work by summarizing our findings and proposing the next logical steps of this research.

CHAPTER II

BACKGROUND

Electronic Medical Record Systems & Security

HCOs increasingly adopt Electronic Medical Record Systems (EMRS) as a means to improve quality while preventing errors (1; 2) and to reduce the cost of delivering care (13; 14). Initially believed to represent a massive step forward in terms of security when compared to the paper-based medical record, EMRSs instead shift the attack vectors (3).

There is an ever increasing body of evidence of privacy violations conducted by insiders who, while authorized to use the EMRS, improperly access patient information. Numerous celebrities, including George Clooney, Britney Spears, and Farrah Fawcett, are victims of improper medical records access by valid EMRS users (15-17). Although these high profile cases garner much media attention, there are likely dozens or hundreds of improper accesses that occur and are never caught. Some improper accesses are likely out of human compassion—wondering how a co-worker is doing during her stay in the hospital—but there are numerous cases of privacy violations for the sole purpose of conducting fraud (18; 19).

A core benefit of an EMRS, centralization, allows providers from distant locations to view and contribute to a unique record for each patient. Centralization means that the records potentially could be protected by strict access control polices compared to paper records, which could lay exposed on a desk or a cart in a hallway. But, centralization also

implies that a breach of privacy can be carried out from a distance and almost instantly.

Many researchers and security officials realize the security implications of an electronic record and propose solutions often based on successes in other industries. Barrows and Clayton stress the importance of healthcare policy informing security policies (3). Their work suggests implementing access control in EMRS, in which each user's permissions to access specific patient information is explicitly defined pre-hoc. Most interestingly, the authors point out the paradox that exists with healthcare data: the most private and sensitive data in a patient's record is often the most relevant data when treating that patient. Denley and Smith have similar sentiments about the use of access control in EMRS (20). They suggest that the access control should be based upon a user's role in the organization (nurse, clerk, junior doctor, radiologist, etc.) and the locations of their responsibility (by ward, specialty, etc.). They indicate several cases in which this model breaks, such as if a user legitimately needs to access the record of a patient who visited the ward beyond a certain time window. Furthermore, Denley indicates that any access control system for medical records needs to have a security override method. These researchers provide sound justification for using access control in healthcare systems, but they fail to address the practical issue of actually building the access control policies.

To a large extent, HCOs successfully adopt many standard information technology security protocols, such as encryption, digital signatures, firewalls, and user authentication. These measures are primarily designed to prevent unauthorized access, rather than to prevent the insider threat by enforcing rules on how users should be

allowed to interact with the data once they have a legitimate need to access the system. Unfortunately, this insider threat proves a greater and more costly challenge to HCOs (7; 21).

The typical method for preventing insider attacks, role-based access control, used by the banking industry and the military, does not work well in the healthcare environment. In role-based access control, each user is assigned to a role. The roles are given the ability to conduct certain tasks within the system. These systems rely on a complete upfront definition of the access policy in terms of the users, roles, and tasks. Due to the highly dynamic and critical nature of healthcare, it is nearly impossible for privacy administrators to manually conduct this upfront or pre-hoc definition of policy within the system. The fear of being too strict in the access policies permeates healthcare. Providers are often unwilling to give up control when providing care to patients, especially in emergency situations.

Government Regulation

Until recently, the most prominent national legislation dealing with healthcare security and privacy was the Health Insurance Portability and Accountability Act, known as HIPAA (22). Within HIPAA, the Privacy Rule details the criteria for "protected health information," how this information can be disclosed, and how improper disclosures are handled. Also within HIPAA is the Security Rule, which indicates general safeguards for protecting healthcare data, including the use of auditing (§164.312(b)) and access control (§ 164.312(a)). Unfortunately, these two rules, while mandating goals, lack specifics of

how these technologies should be implemented. This lack of specificity puts the healthcare industry in a state of limbo—in which the industry acknowledges that its data must be secured, but lacks full, end-to-end security and privacy solutions.

While HCOs in the United States work to become HIPAA compliant, there have been few, if any, sanctions imposed for security or privacy lapses. But recently, government agencies have started pushing HCOs to comply with the legal codes by levying heavy sanctions (23).

In response to the ambiguous regulations of the past decade, the recent passage of the Health Information Technology for Economic and Clinical Health Act, part of the American Recovery and Reinvestment Act of 2009, is expected to improve the regulation of HCO security (24). As part of the act, HCOs must simultaneously improve adoption and security of electronic medical record systems. The act seeks to improve security by forcing HCOs to monitor and publicly report any breaches of patient privacy. This renewed regulatory interest in healthcare information security justifies the need for the work that we are presenting.

Log Mining

A fairly young, but rich history exists for mining system access logs. Much of the work in log mining parallels the growth of the Internet. Some of the early work in access mining comes from Cooley, Srivastava, and colleagues. They discuss pattern discovery from web server access logs (25; 26). They additionally hint at the use of association rule mining as a method for performing the pattern discovery. They also discuss methods for

preparing data, specifically with the intent of tracking user sessions (27).

Based on these techniques, researchers have begun to successfully use log mining to discover patterns in human behavior. From email logs of who sent and received email within Enron during the company's collapse, researchers can detect fundamental organizational changes (8; 28).

In the healthcare setting, access log mining is often conducted in connection with improving electronic education resources (29-31). The work of Chen et al. shows that mining EHRS access logs is useful for discovering clinicians' information needs (32). Additionally, Malin's research shows that patients' Internet browsing patterns on a health information site are correlated to those patients' medical diagnoses (9).

The work of these researchers indicates the usefulness of access logs as a source of information about how users and organizations interact and change. Our proposed research extends the field of log mining by applying this field's techniques for the purpose of generating security policies.

Process Mining & Workflow Analysis

The concept of process mining or workflow analysis in healthcare is explored by several groups (33). The study of workflows is especially important to healthcare because it can help detect and fix suboptimal processes in an effort to increase the quality of care (1). The cases of HCO workflow mining are often localized to a small group and focused on safety improvement measures, instead of HCO-wide privacy and security issues.

Outside healthcare, the study of workflows is also active, from the

characterizations of how workflows and organizations change over time (34) to advanced techniques to detect, model, and analyze workflows (35-37). Unfortunately, many of these techniques only work for small datasets, and therefore are not suitable when analyzing a whole organization.

Conclusion

Our work comes at an opportune time. There is an existing void in healthcare security, caused by the inability of traditional solutions to prevent insider attacks within HCOs. Unfortunately, this void is increasingly exploited as EMRSs become more prevalent. With the public aware of this growing privacy issue, the government is actively pressing HCOs to close the security void. In the coming chapters we will present a framework takes the first step towards addressing the current security void.

CHAPTER III

HORNET SOFTWARE ARCHITECTURE

Introduction

One of the main contributions of this work is an open-source software platform (HORNET) for efficiently analyzing HCO access logs. The design of HORNET includes two primary goals: reusability and extensibility.

In order to be reusable, HORNET uses abstract concepts and allows administrators to configure mappings from the organization-specific details into these abstract concepts. For example, specifics of the access log format, such as delimiters and fields, are configured by the user. Configuration of HORNET allows other HCO administrators and researchers to reuse HORNET to conduct the analysis presented in this thesis on their own information systems. In order to analyze small to large HCOs, HORNET must also perform its analysis efficiently.

In addition to being reusable, HORNET is extensible, so that developers can add or customize analysis techniques. HORNET utilizes a plugin architecture, in which developers can take advantage of a rich Application Programming Interface (API). This API allows quick development for additional features and analysis techniques. Since HORNET is an open-source project, released under the Apache License, Version 2.0 (38), developers can safely extend and contribute to the platform.

HORNET Within An Existing HCO Infrastructure

To facilitate the adoption of HORNET by other HCOs, we have designed it to seamlessly fit into the existing IT infrastructure of an HCO. HORNET will accept access logs in nearly any format from any source IT system. The access logs can be in a text format, such as comma- or tab-separated or XML. Likewise, the access logs can be stored in any major relational database system, such as Oracle, MySQL, Postgres, DB2, Access, SQL Server, or SQLite. The ability to accept multiple data formats comes from our abstracted model of an access log, with specific implementations that map XML, flat text, or relational database tables into abstract access events. Once HORNET is installed and configured to the format of the access logs, the user can select which analyses he or she wishes to run, using a configuration language. As shown in Figure 1, HORNET sits on top of the existing HCO information technology infrastructure, thus no changes beyond obtaining log files is necessary for HORNET. HORNET can optionally feed off of sources of meta-information about users—such as a user's role or department that is stored in traditional human resource databases.

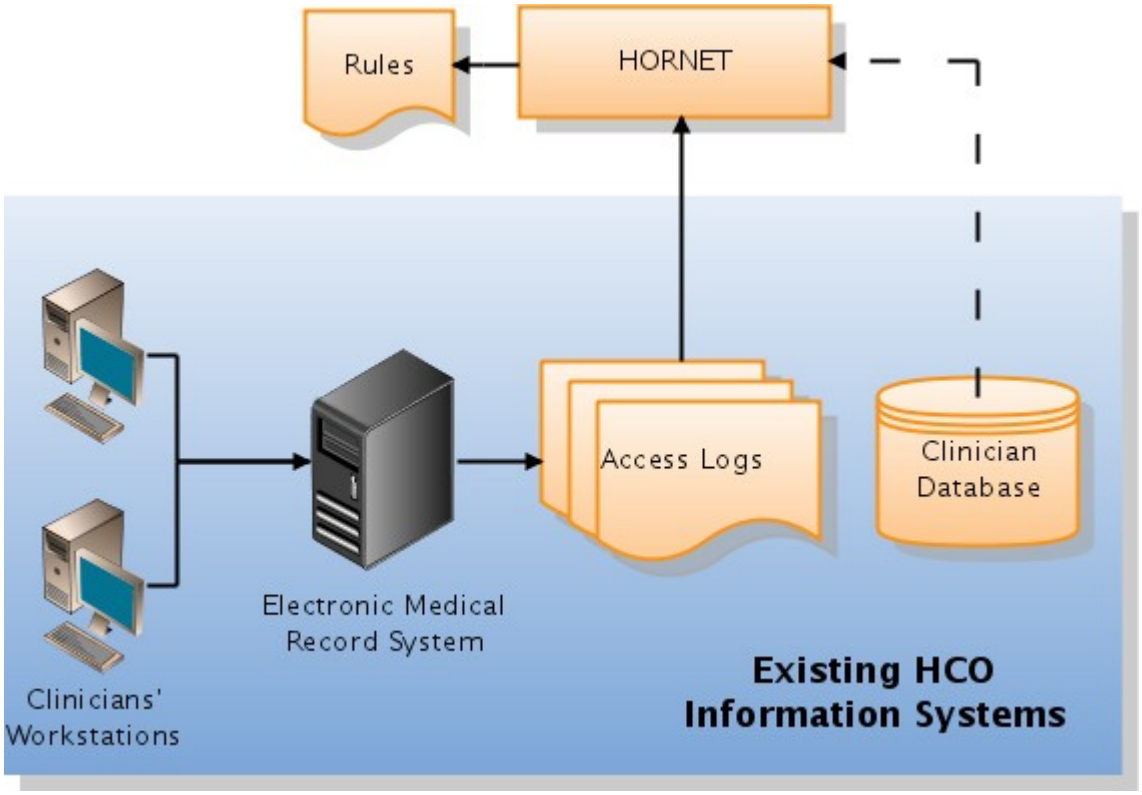

Figure 1: HORNET requires no changes to the existing HCO information system infrastructure.

Software Description

The HORNET software platform, developed in the Python language (39) has two key pieces: HORNET Core and the HORNET plugins (Figure 2). The HORNET plugins each perform a specific task, usually involving some specialized method of analysis, while HORNET Core is a general framework on which the plugins run. The Core additionally provides common functionality (e.g., file and database access), general data structures (e.g., graphs, nodes, and edges), and plugin configuration. New plugins can be written for alternate analysis and use HORNET Core to remove much of the "grunt" work.

HORNET Core consists of a handful of modules that ease plugin development, shown in Table 1. HORNET Core also provides the "executable" that launches any desired analysis as well as the configuration mechanism to allow the user to specify what analysis is desired. HORNET is capable of running on any modern operating system.

Table 1: HORNET Core Modules

Module	Features
File API	• Read/write delimited files, such as comma- or tab-separated data • Create unique, temporary files • Move & delete files • Persist complex data structures to disk
Database API	• Query major relational database products (including MySQL, Postgres, Oracle, SQLite, Access, SQL Server, and DB2) • Wraps the functionality provided by the open-source SQLAlchemy[1] library
Network API	• Basic data structures, such as a relational network, edge, and node • Methods to manipulated and analyze these data structures • Wraps the functionality provided by the open-source library NetworkX[2]
Plotting API	• Create common graphs, such as XY-plots or log-log plots • Wraps the functionality provided by matplotlib[3] library
Task API	• Run computationally intensive tasks in parallel, by taking advantage of multiple processors and multiple cores
Plugin API	• Specification to which plugins must adhere • Provides common utilities, such as logging

1 http://www.sqlalchemy.org/
2 http://networkx.lanl.gov/
3 http://matplotlib.sourceforge.net/

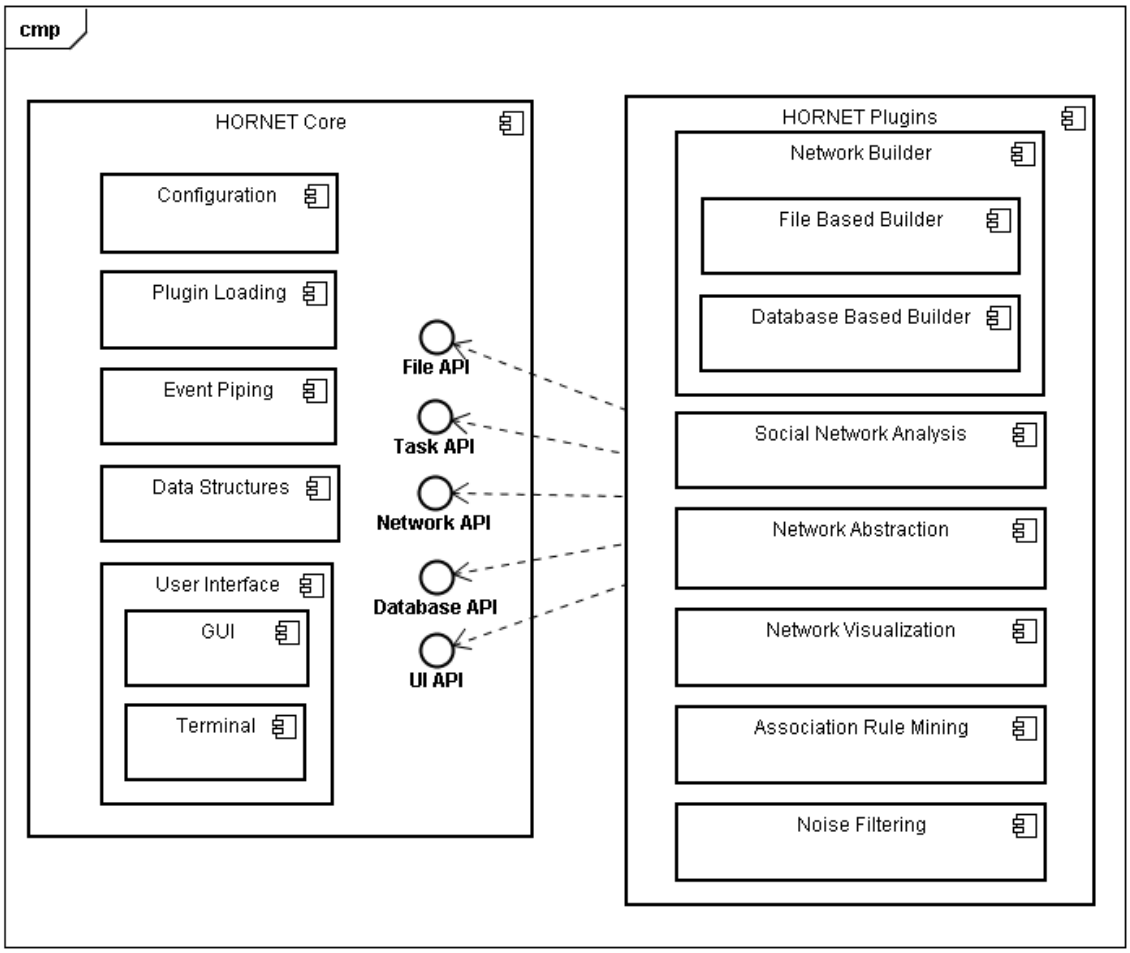

Figure 2: The modular design of HORNET allows easy extension.

Plugin Architecture

On top of HORNET Core, we provide several plugins and allow for any additional plugins (Figure 2). The plugins must adhere to a basic contract which defines how they are notified that it is their turn to process data and how to return data once they are finished with their analysis. Beyond this basic contract, the plugins can perform any type of operation a developer desires. The basic contract allows for plugin "chaining" in which the output of one plugin becomes the input of another plugin. For example, one

plugin can create a relational network object from a log file, HORNET Core then passes this relational network object to the plugins that are configured to wait for this data. As Figure 3 shows, the network object is passed to the Social Network Analysis plugin and the Network Abstraction plugin. Once these plugins finish their analysis, their output is appropriately piped to the next plugins. The plugin chaining can become complex, allowing for forks in processing (i.e., multiple plugins can process the output of a single plugin) and for joining forks (i.e., forked execution threads can be rejoined into a single plugin). The plugins and the piping configuration are specified in a single configuration file, using a powerful configuration language, which is a subset of the Python language.

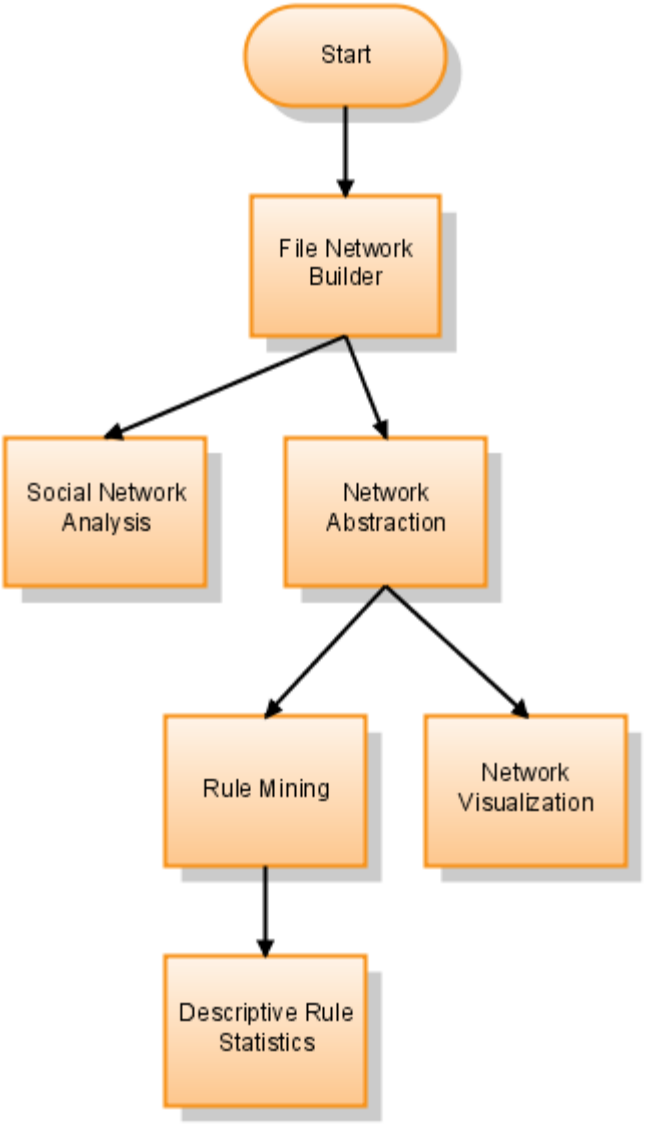

Figure 3: Example of HORNET's Plugin Chaining

Runtime Information

HORNET is designed to run efficiently, providing administrators or researchers results within a reasonable time frame. Since HORNET sits on top of the existing infrastructure, it can process data in an off-line, batch mode.

Many of the analyses that HORNET performs are CPU-intensive. However,

HORNET can be memory-bound depending on the size of the dataset. The memory required to store the data structures scales linearly with the number of relationships between users.

We use several techniques to improve the runtime of HORNET. First, the data structures and algorithms are optimized to reduce the size of objects in memory and the number of computations required (further optimization is very likely possible). Second, the Task API allows parallel computation. We use the Task API to analyze multiple time periods in parallel. Third, if memory is an issue, HORNET can be configured to run in several steps, in which intermediate results are persisted to disk.

As a point of reference, the retrospective study presented in Chapter V, which analyzes 21 weeks of access logs from VUMC, takes less than 12 hours on a machine with a dual core processor and 4 gigabytes of memory.

Development Information

HORNET is a command-line application, which is programmed in Python, and currently optimized for Python version 2.6. The application consists of over 6,600 lines of source code, more than half of which are unit tests and functional tests to ensure the validity of the logic. Over 93% of all code statements are unit tested. The code contains in-file documentation, including sample use of the code, to assist plugin developers.

As an open-source program, we provide a public, version-controlled, source repository on Google Code[4]. Third-party developers are encouraged to submit changes

4 http://code.google.com/p/hornet/source/browse/

using the Mercurial version-control tool. We also host an issue tracker on Google Code[5], to which we encourage the submission of bug reports and feature requests.

Upon every commit to the source code repository, our continuous integration server executes all the unit and functional tests and rebuilds the searchable documentation[6]. This documentation is built from the in-file documentation and examples as well as from supplementary files detailing how to use, install, and develop HORNET (Figure 4). If a commit causes a test to fail, the developers are notified via email, so that a prompt fix can occur.

We selected a permissive open-source license, the Apache License 2.0, which allows any HCO to use and extend HORNET.

Future Plans

HORNET's open nature, both in terms of its licensing and its plugin architecture, means that other developers and researchers can be part of the future direction of HORNET. We specifically plan several major enhancements, including a graphical user interface. We are also in active development of other plugins. These plugins will experiment with different units of analysis and different methods of analysis. Additionally, we plan to develop a plugin that uses the statistical model we present in the following chapters to characterize if the accesses of individual users are suspicious.

5 http://code.google.com/p/hornet/issues/list
6 http://hiplab.mc.vanderbilt.edu/projects/hornet/snapshot/docs/

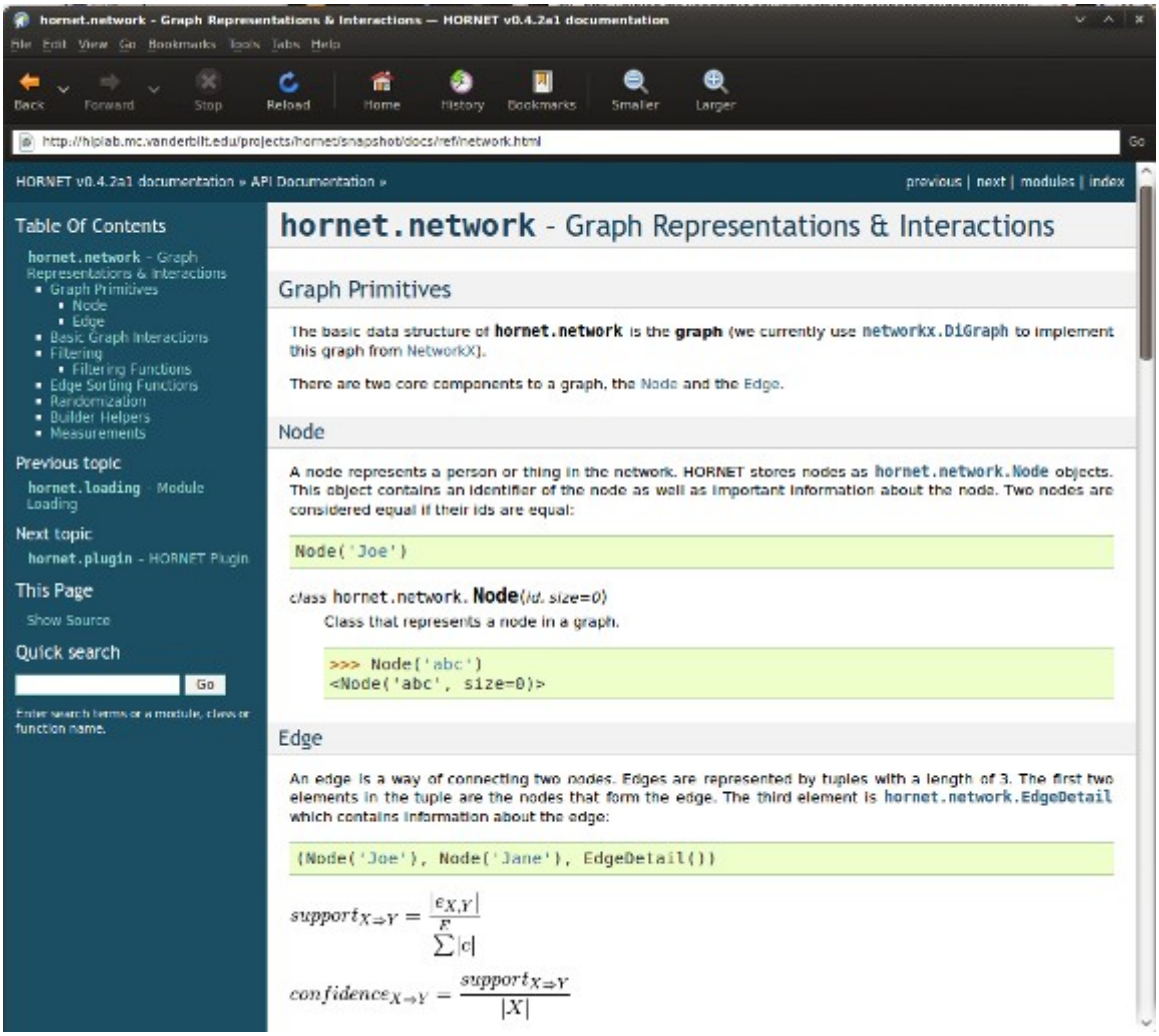

Figure 4: Example of HORNET API Documentation

CHAPTER IV

STATISTICAL MODEL OF THE ORGANIZATION FROM ACCESS LOGS

Model Description

Our goal is to obtain a statistical model, which describes the interactions and relationships of providers within the HCO based upon their common interactions with patients, from the access logs of an EMRS. This model characterizes how the organization works by using the EMRS as a proxy for normal processes and operations within the HCO. We will first discuss the basics concepts of our relational network model, then cover how to build this model from access logs. We use the HORNET framework discussed in the previous chapter to create a plugin that builds this statistical model.

Modeling the Organization as a Relational Network

The techniques we present are rooted in the social networking community, which has a rich history of modeling and analyzing bodies of people (40). This history includes the famous small world experiments of Milgram and Traver that shows that any two people in the United States are separated, on average, by only 6 other people (41; 42). Recent re-creations of Milgram's study show similar results for the connections of users of instant messenger systems (43) and blog links (44). In the field of Biomedical Informatics, social network techniques are valuable in understanding the structure of

editorial boards of journals in the field (45) and in characterizing the focus of departmental research interests (46).

The basic unit of analysis of social networking is a user, also called a node, and represented as *n*. Two nodes can be connected by an edge, labeled *e*. The set of all nodes, *N*, and set of all edges, *E*, exist within the network (also called a graph), *G*. In *G*, an edge, $e_{x,y}$, exists between two nodes, *x* and *y*, if those users accessed at least one patient's record in common. We denote the distinct number of patients a user has accessed by $|n|$ and the number of patients two users accessed in common by $|e_{x,y}|$. Finally, we keep track of how many distinct patients in total exist as $|G|$.

We can examine a trivial example, as shown in Figures 5 and 6. Imagine there are three users, who access a total of four patients' records. We initially show these two classes of people in a traditional bipartite graph (Figure 5). Our goal is to transform the bipartite graph into a relational graph (Figure 6), in which the patients become a property of the relationships or edges between users. User A is the only person to access Patient 1, so $|User\ A|=1$. User B accesses Patients 2, 3, and 4, thus $|User\ B|=3$. User C accesses Patients 3 and 4, thus $|User\ C|=2$. The total number of patients in the network, $|G|$, is 4. Since Users B and C accessed two patients in common, there is an edge between their two nodes, with $|e_{User\ B,\ User\ C}|=2$.

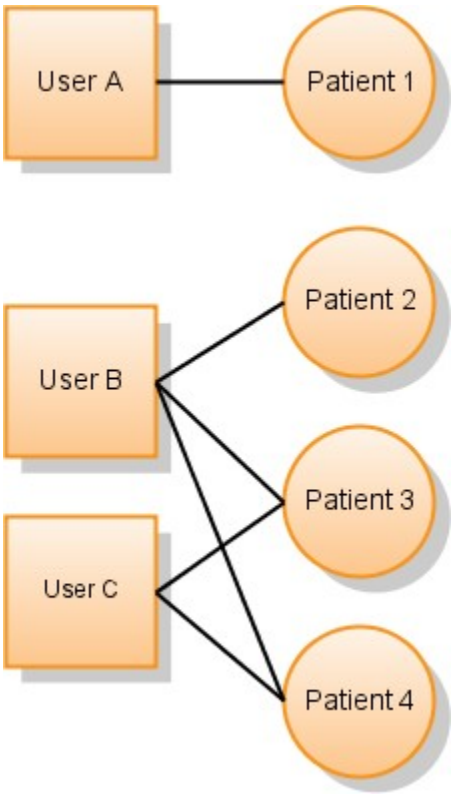

Figure 5: Simple Example of Bipartite Graph

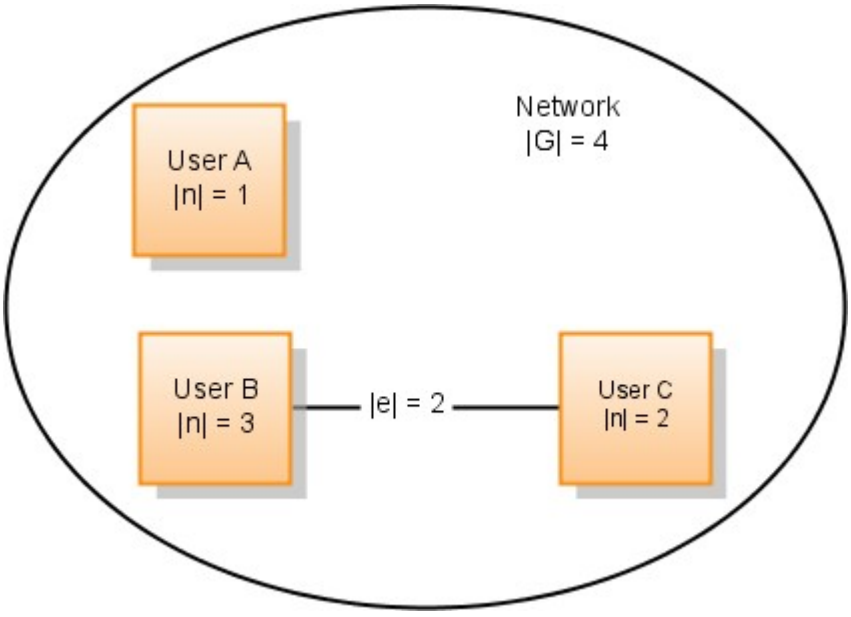

Figure 6: Bipartite Graph Converted to Relational Network

From this rather simple relational network construct, we can compute several interesting metrics and statistics. Using a similarity metric, such as the cosine distance (47), we can quantify the strength of relationship between two edges, with respect to their other relationships. We compute the cosine similarity as:

$$similarity = \frac{|e_{x,y}|}{\sqrt{|x|^2}\sqrt{|y|^2}}$$

Thus, for each node, we can rank-order that node's edges, essentially providing us a list of strength of "friendship." It is important to note that when we use the words "friendship" or "relationship" we mean purely from the perspective of the relational network. It is very likely that in the actual organization two people who are in a common care pathway may never know each other, yet they are connected from our perspective because they care for the same patients.

In addition to similarity, we can compute several probabilities related to the user accessing a record. We refer to the probability of a patient's record being accessed by two specific users, such as *x* and *y*, as the pair probability, and compute it as:

$$P(x,y) = \frac{|e_{x,y}|}{|G|}$$

The conditional probability is the probability of a patient's record being accessed by a certain user, *y*, given that the record is accessed by a certain different user, *x*:

$$P(y|x) = \frac{P(x,y)}{|x|/|G|} = \frac{|e_{x,y}|}{|x|}$$

In the conditional probability equation, we refer to *x* as the antecedent and *y* as the consequent. We sometimes refer to the relationship between *x* and *y* as the rule between *x* and *y*.

Extracting the Relational Network from Access Logs

In order to create this relational network of the HCO, we need system access logs. Nearly all Health Information Systems generate access logs for the purpose of generating an auditing trail of what information providers see and what actions they take. The access logs may be stored as plain text, XML, or within a database. We can build our model using either files or a database with the same approach, but for simplicity we will just refer to them as files. Typically, each line of the file represents a single action in time by a user as demonstrated in Figure 7. A line should contain at the very least a timestamp of when the action occurred, a unique identifier of the user performing the action, and a unique identifier for the patient. The line may have additional information that is relevant to the EMRS, such as the client's IP address and the action taken by the user.

```
Timestamp           User ID      Patient ID   Client IP     Page Accessed
01/01/2006 02:23:47 drsmith      012345       10.127.0.1    view.cgi
01/01/2006 02:24:02 drsmith      999999       10.127.0.1    view.cgi
01/01/2006 02:24:03 drjones      012345       10.127.0.2    view.cgi
01/01/2006 02:38:00 drsmith      012345       10.127.0.1    view.cgi
...
```

Figure 7: Example Access Log

With the timestamp, user identifier, and patient identifier from the access logs, we construct the relational network according to the steps in Table 2. If desired, we can partition the access log entries based upon some time-period, such as 7-days, and create individual networks for each time period.

Table 2: Steps to Generate Relational Network

Step	
1	Create an empty network, G.
2	Find the number of distinct patients, set this number into $\|G\|$.
3	For each user, n, create a list of patients that user accessed. $\|n\|$ = size of this list. Insert n into G.
4	For each possible pair of users, x and y, find the intersection of their lists. $\|e_{x,y}\|$ = size of this intersection. In G, connect x and y with an edge, $e_{x,y}$.

Statistics from the Relational Network

After using the steps in Table 2 to generate a relational network, we can easily compute the pair and conditional probabilities using the equations we previously defined. A high pair probability for this edge would indicate that this is a very common relationship within the HCO, such that a large number of patients are accessed by both providers. High pair probabilities indicate that the users are likely in high volume care areas. However, given the typical delivery of care by all but the smallest HCOs, it is unlikely that any set of users will have relatively large pair probabilities. For example, in order for an edge to have a pair probability of 0.10, both users must access at least 1 in every 10 patients.

While the pair probability is useful for putting each relationship in perspective to all the other relationships, the conditional probability gives a more local view of the relationships and allows us to find strong relationships that may by rare according to the pair probability. For example, two users may be part of an workflow in a care area that only sees a handful of patients in a week. The pair probability of this edge will be

dwarfed by other edges representing groups such as the emergency department who likely see dozens to hundreds of patients a day.

This relational network and statistical model is simple, yet powerful for reducing the huge amount of data that access logs provide into likelihoods of seeing certain events within the EMRS.

Abstracting the Low-Level Network into a Generalized Network

While examining how specific providers interact within the system is informative, this provider-level model can be very brittle when modeling modern HCOs due to the way in which providers act in roles and teams. For instance, if a patient is being treated in an Intensive Care Unit, it may be more informative to know that a critical care nurse accessed the record than to know the particular identity of the nurse. Understanding who the providers are at an abstracted level empowers our model to make broader generalizations of the HCO.

Additionally, these abstractions could potentially smooth volatility from the model. For example, several nurses and fellows may perform the same actions on the same set of patients but work at different times of the day. At the user level, we will detect different clusters of relationships depending on the time of day. But if we abstract the data appropriately, we could remove the time factor to study true care roles.

We can construct abstracted versions of the relational network if, in addition to the access logs, we have a mapping of the users to some generalization, such as their department (Critical Care, Medicine, Oncology, etc.) or their role (fellow, nurse, biller).

We refer to this type of information about users as "meta-information." The abstraction process (Figure 8) takes in a relational network and returns a new relational network where nodes are at the generalized level and the edges between these generalized nodes represent the combined edges of the constituent user-level nodes. If the generalization mappings indicate that a user has more than one generalization (e.g. an attending physician has appointments to two distinct departments), we discount the user's contribution to the abstracted edges and nodes proportional to the number of generalizations for that user.

```
function m(n)
        return the list of meta-information mappings of n

Create new network, G', with |G'| = |G|.
For each node, n, in G,  find m(n)
        If m(n) is empty, ignore n and continue
        Else for each n' in m(n)
                If n' already exists in G', add |n| / size(m(n)) to |n'|
                Else insert n' into G', with |n'|=|n| / size(m(n))
For edge, e_x,y, in G, find m(x) and m(y)
        If m(x) or m(y) is empty, ignore e and continue
        Else for each x' in m(x) and y' in m(y)
                If e_x',y' already exists in G', add |e_x,y| / (size(m(x)) * size(m(x))) to |e_x',y'|
                Else insert e_x',y' into G', with |e_x',y'|=|e_x,y| / (size(m(x)) * size(m(x)))
```

Figure 8: Pseudocode for Abstraction a Relational Network

As an example in performing an abstraction, we will use the sample relational network in Figure 6. Let us define the meta-information mapping such that User A has no mapping, User B maps to Dept 1 and Dept 2, and User C maps to Dept 3. Using the algorithm in Figure 8, we discount the nodes and edges of User B since User B has two

department abstractions, as shown in Figure 9.

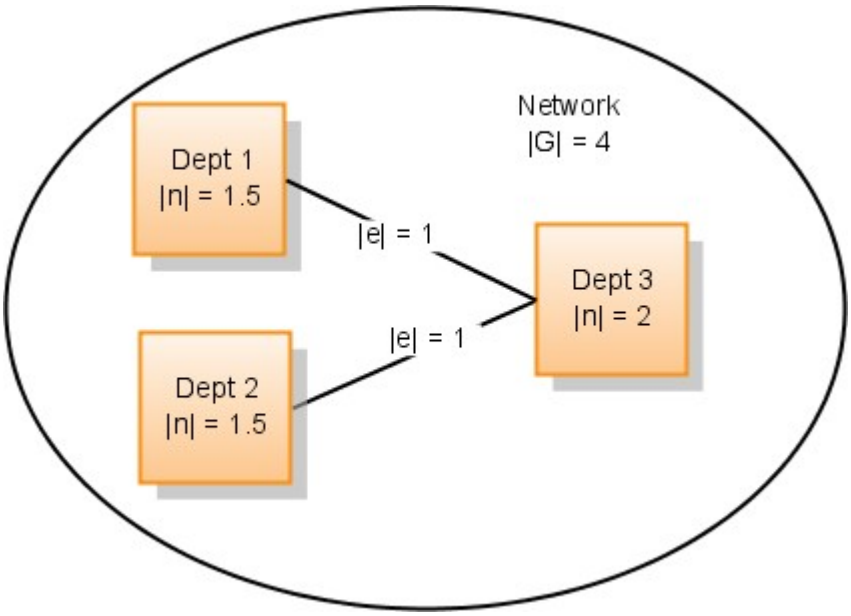

Figure 9: Sample of an Abstracted Relational Network

Implementing as HORNET Plugins

We implement this approach using HORNET's plugin framework. To build the network graph, we use a File or Database Network Builder plugin, which reads the access logs. As it inserts nodes and edges into the network, it begins calculating the pair and conditional probabilities of each edge. Once complete, HORNET pipes the resulting network objects to any plugins configured to receive the Network Builder's results.

One such receiver is the Network Abstraction plugin, which connects to a pre-configured data source of meta-information. Using this data source, the plugin joins nodes that have common meta-information (e.g., the same department). When the join occurs, the plugin appropriately joins any associated edges and re-computes the edge

probabilities. Once the abstraction is complete, HORNET passes the abstracted network object onto any plugins configured to receive its output.

CHAPTER V

USAGE OF HORNET ON STARPANEL ACCESS LOGS

In Chapters III & IV we introduced a framework for mining healthcare access logs then, using this framework, built specific methods for representing and analyzing the data in these access logs. In this chapter, we demonstrate these techniques by analyzing access logs from Vanderbilt University Medical Center's (VUMC) EMRS. Specifically, we present results on how users and departments interact within VUMC, and we demonstrate how inherent organizational structure makes examining the organization in an abstracted manner more appropriate. Additionally, we discuss issues in using data from real-world information systems such as incomplete data.

Materials

We wish to use the plugins presented in Chapter IV to build a relational network that characterizes the probabilities of clinicians caring for the same patient using data from VUMC's EMRS, StarPanel. VUMC is a large, tertiary care hospital located in Nashville, Tennessee.

VUMC has developed a custom longitudinal medical record system, StarPanel (48-50), based upon the MARS clinical repository, originally developed at the University of Pittsburgh (51). StarPanel is a web-based system, which employs a number of servers that users access through a web browser. These servers track usage of the system by

writing a summary of each action a user takes within the system to a log file. These log files provide an auditing capability, however, as discussed in Chapter II, the task of manually auditing each action is monumental. Each audit event stored in these logs consists of a timestamp, the user's information (username and Internet Protocol address), and information on the resource the user accessed (e.g., server file and patient medical record number).

We obtained a 21-week sample of the StarPanel access logs, dating from January 1st, 2006, to May 27th, 2006. These logs are consistently de-identified with pseudonyms, such that each instance of a unique username, IP address, and medical record number is globally replaced with a random ID number. This approach to de-identification removes all identifiable information, while preserving the patterns and context within the log files. In this 21-week sample of data, 9,940 distinct StarPanel users accessed the records of 350,889 distinct patients, resulting in a total of 7,575,434 accesses of patient information.

In addition to the StarPanel access logs, we sought to obtain meta-information about the StarPanel users—information such as the user's role (e.g., physician, nurse, etc.) and the user's department (e.g., Internal Medicine, Radiology, etc.). Several sources each provide some subset of this meta-information, but none represent a consistent, universal dataset that provides information about the users. One source, StarPanel, maintains simple role information for customization of StarPanel based upon a user's self-identified role. Each user has one of seven roles assigned, shown in Table 3. Unfortunately, a large number of users lack a specific role (they are given the role of "??"). This lack of universal identification of role within StarPanel does not represent an

issue with StarPanel, but rather an issue caused by trying to use the data for a secondary, unintended use. A second source of meta-information is the several distinct databases Vanderbilt maintains about clinicians. Information about medical students, residents, and attending physicians is stored in separate databases. In addition to the information being in separate physical databases, the information within these databases is not consistent. For instance, the labels for the same departments were often completely different between databases. The rationale for keeping this information separate is valid, since these populations have intrinsic differences, but it does create issues when trying to use the information for a secondary purpose as we intend in this study.

Table 3: Self-Assigned Roles for StarPanel Users

StarPanel Role	Number of Users
RN	3471
??	2521
MD	1089
AT	945
AN	812
MR	451
MS	377

Ultimately, we use a single source of this meta-information, a database of physicians, nurses, and administrative users, located in VUMC's Enterprise Data Warehouse (EDW). From this database, we link the department of a user to that user's accesses. This database contains department information on 3,687 users, 36.9% of the 9,940 distinct users during our study period. We will discuss the impact and implications

of using only this single provider database later.

We configure HORNET to connect to an Oracle 10g database, which stores the de-identified access logs and this departmental information. We import the access logs into a database as opposed to building the relational model straight from the raw text log files to improve our ability to query the data. Once configured, HORNET processes the access log data using the plugin described in Chapter IV. Specifically, we configure HORNET to detect changes over time by analyzing the data in 21 one-week segments, split on midnight on Saturday. We determined that one week segments were appropriate due to the periodic nature of system usage as shown in Figure 10.

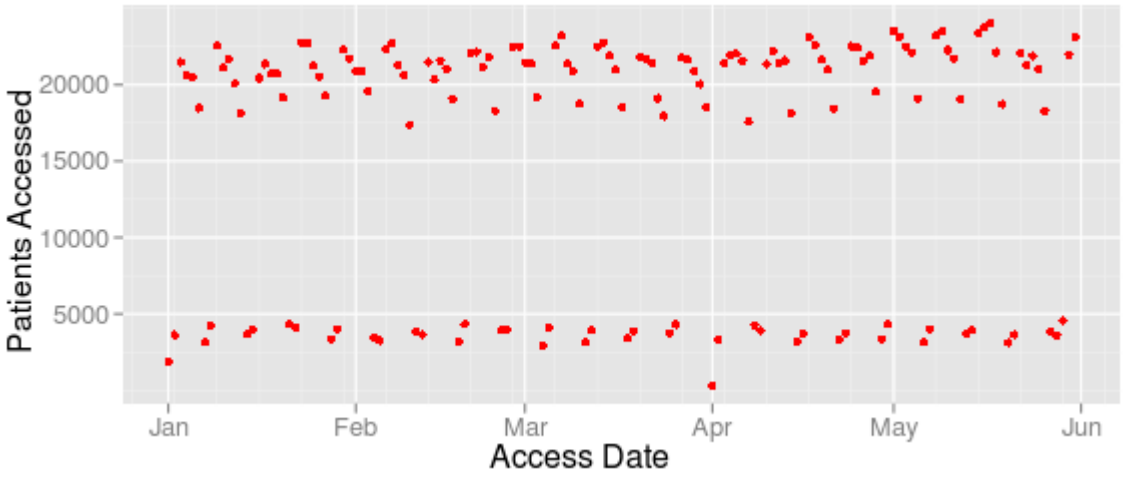

Figure 10: Number of Accesses Across Time. A periodic trend shows much greater system usage during weekdays as opposed to weekends.

For each week, HORNET builds a relational network using the access log data from that week. Using HORNET's plugin piping system, we route the results of the Network Builder to the Network Abstraction plugin, so that the user-user relational

network is abstracted into a department-department relational network, using the department information from the EDW database. We then route the output of the Network Builder and Network Abstractor plugins to several plugins that compute various general statistics on the relational network. These general statistics include the number of relationships in the network.

Ignoring Users When Abstracting Relational Network

When performing the abstraction of the network, HORNET ignores the 63.1% of users who lack department information. We will discuss the implications of the removal later, but we believe we are justified: the two populations (those with departmental information and those without departmental information) have statistically indistinguishable patterns of accessing patient information. A Mann-Whitney U test rejected the null hypothesis that the two groups are different in terms of how many records each user in the two groups accessed, with a $p < 0.001$. Additionally, when we compare the a complete version of the user-level network to a version with all users who lack department information filtered out, we see that the ranks of the remaining rules do not change in relational to each other. So removing the users still keeps a meaningful sub-network.

Results

Using HORNET to transform the access logs into a relational network allows us to examine some basic properties of the way in which Vanderbilt clinicians operate. We can see that the number of users accessing patients' records in StarPanel is consistent

throughout the 21-week study period (Figure 11), with an average of 6,406 users per week and a standard deviation of 126. Relatedly, the average number of departments each week is 292 with a standard deviation of 4.5.

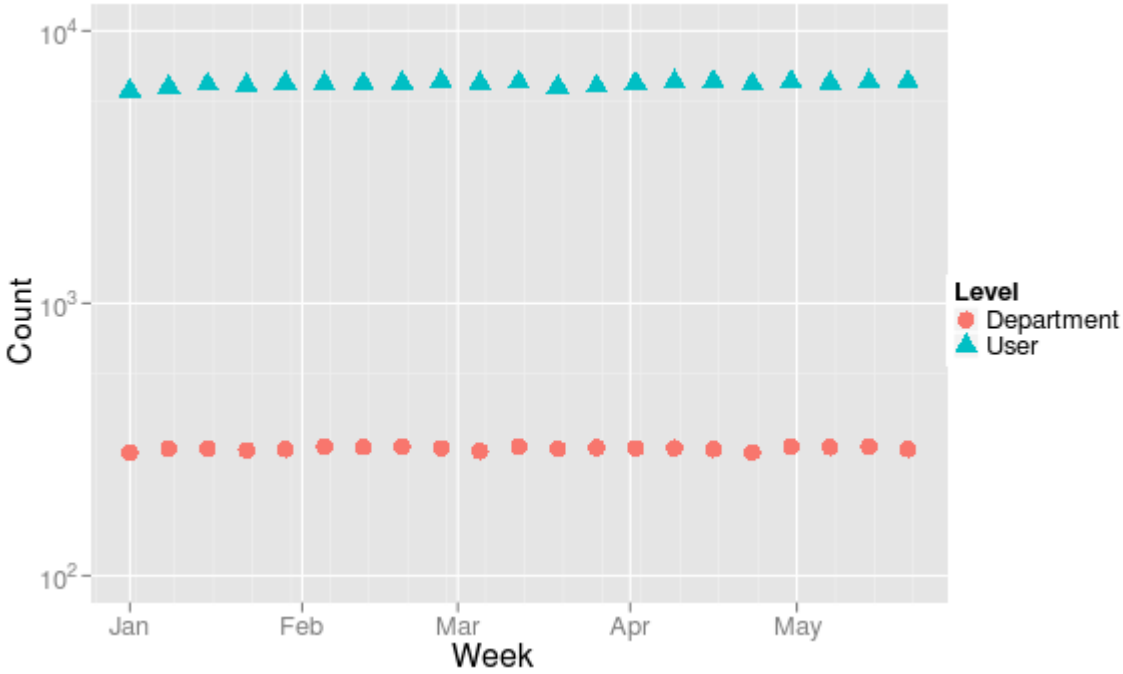

Figure 11: Number of Departments and Users each week.

While the system usage is consistent, we see a huge variation in the number of patients' medical records a clinician accesses, as Figure 12 indicates. If we look at the week of April 23, 2006, arbitrarily picked from our 21-week sample, we can see that 861 of the 6,389 users accessed only a single patient's medical record, while one user accessed 1,097 records. The number of records accessed by a user in one week on average is 28.8, with a median of 11. The distribution of the number of patients accessed by each user appears to follow a power-law distribution, a property that is often seen in

social networks.

This distribution indicates that a large number of users interact with only a few patients' medical records each week, while a small number of StarPanel "power users" access hundreds of patients' records. We can confirm this concept of "power users" by examining the job titles of those users who were on the upper end of this distribution. The user who accessed 1097 patients' records is a medical records analyst in the Emergency Medicine department. The person who accessed the second most patients in the week of April 23rd, is a coding specialist. Both of these users have jobs that entail accessing the medical records of numerous patients.

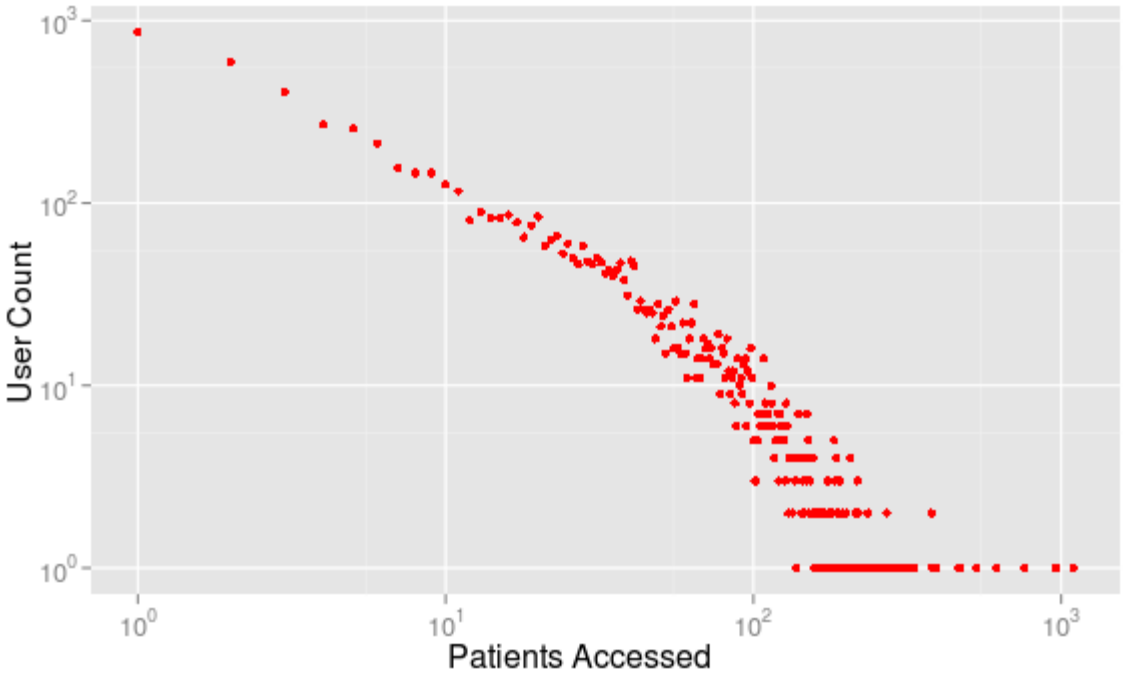

Figure 12: Patients accessed per user for the week starting April 23, 2006. All other weeks during the 21-week study period had a similar power law distribution.

Interaction Probabilities

Moving beyond these simple network statistics, we can look at the rules, which form the basic probability model for the likelihood of users and departments accessing common patients. As Figure 13 shows, the number of rules detected each week is extremely consistent. In terms of user-user probabilities, we detect on average 886,784 rules per week, with a standard deviation of 48,972. There is an upper limit to how many of these user-user relationships are possible, given by $\frac{n^2-n}{2}$, where *n* is the number of users. Using the average number of users per week, 6406, we find an upper limit of 20,515,215. With only 4.32% of all possible relationships, we see that the user-level network is relatively sparse—indicating the possibility of community structure withing the organization.

If we look at the abstraction of these user-user rules into department-department rules, we detect 27,261 rules on average per week with a standard deviation of 862. There is also an upper limit to the number of department-department relationships, given by $\frac{n^2-n}{2}+n$, where *n* is the number of departments. We add an additional *n* in this equation because it is possible for departments to have relationships to themselves. Using the average number of departments per week, 292, the upper limit is 42,778. We discover 63.7% of the possible department relationships, indicating a greater amount of cross-community relationships at the department-level than we see at the user-level.

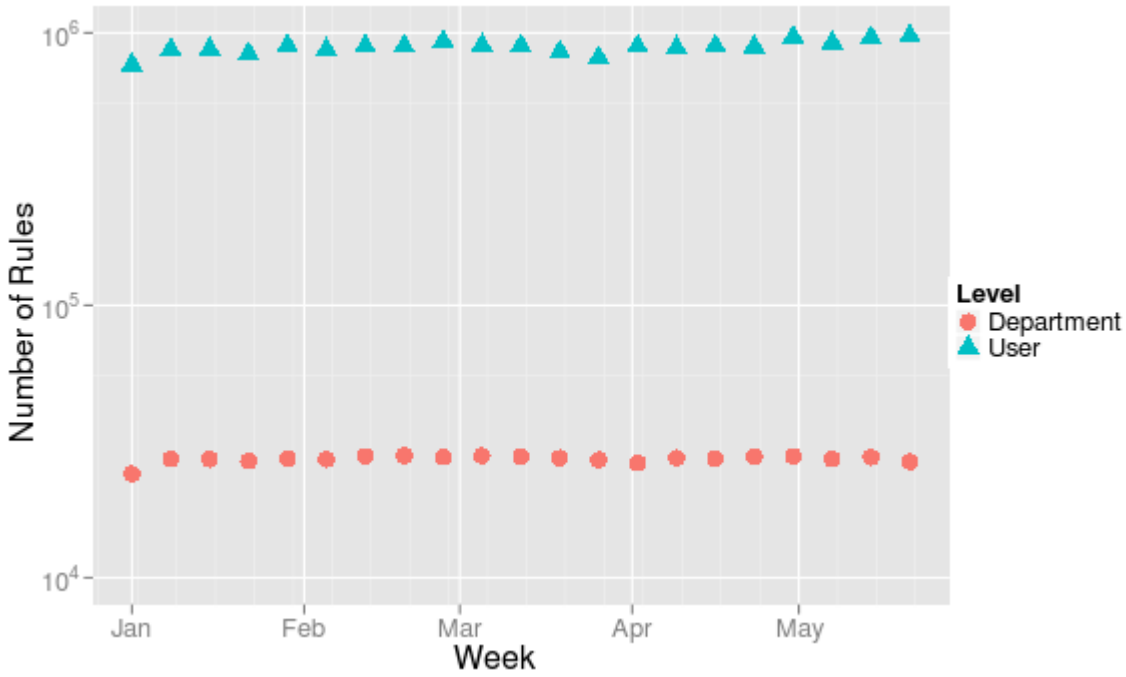

Figure 13: Number of rules each week at the User and Department levels.

When we examine the distribution of relationships per user for the week of April 23rd (Figure 14), we discover that it roughly follows a power law distribution. This highly skewed distribution has an average of 139 rules per user and a median of 77. The distribution shows that there are certain users of StarPanel who are the only person accessing a patient's record in that week, while other users are connected to a very large number of other users. In the week of April 23rd, 118 users (1.8% of the users during the week) were the only users accessing a specific patient's record, and 95 users (1.5% of user) only accessed records that one other person accessed. At the other end of the spectrum, one highly connected person had rules with 2,584 other users.

When we look at the same measurement at the department level, the number of rules per department in a week, we do not find any clear distribution. Nonetheless, the

average number of relationships for each department was 98, with a median of 88.

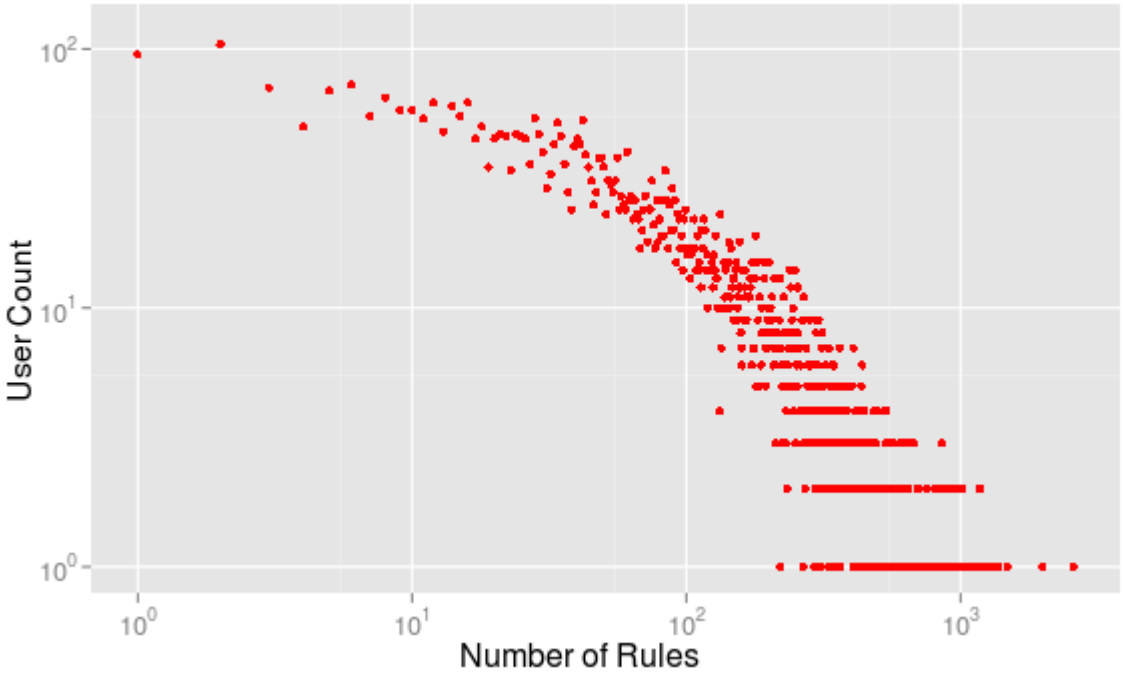

Figure 14: Rules per user for the week starting April 23, 2006. This distribution is consistent with the distributions seen for the other 20 weeks in our study period.

We can speculate the nature of types of accesses for both ends of this distribution. For the records that were accessed a single time, the user may have been accessing the record for a retrospective chart review or renewing a patient's prescription. Alternatively, the low access count could be an artifact of how we split the data into weeks; for example, the physician could be writing a clinical note on a Sunday when the patient had been discharged on Saturday. In future work, we can investigate if splitting the data on a weekly is the cause by exploring larger time periods. Patients who receive a much larger number of accesses are most likely actively being cared for by multiple clinicians.

Rule Decay

We now know the basic distributions of how many patients a user or department accesses, as well as how many rules those users and departments are involved in, but these distributions tell us nothing about how the relational network evolves over time. At one extreme, we can imagine an HCO that has a static relational network due to extremely consistent behavior of all parties over time. For example, a theoretical practice with 2 doctors and 3 nurses will likely be extremely stable over time as long as no one leaves the practice, since on average the 3 nurses will likely interact with the 2 doctors consistently, and the 2 doctors will likely have independent sets of patients they treat. At the other extreme, we can imagine an HCO that has an extremely dynamic relational network. The dynamic nature of this HCO could be caused by having a large number of clinicians and by having multiple clinicians who can all perform roughly the same job. Since multiple people are essentially interchangeable, the possible combinations of care paths increases dramatically. It is important to note that we are not suggesting that clinicians are commodities, but rather that as the number of qualified clinicians in the organization increases, the probability of a patient being assigned to a specific clinician decreases.

To learn how static or dynamic VUMC is, we wish to find the number of weeks a given rule exists during our 21-week study period. A rule that appears for all 21 weeks shows that those two users or two departments accessed records in common each week. The more rules that exist for all 21 weeks, the more evidence we have for a stable organization structure. If rules exist for only a few of the 21 weeks, we have more

evidence that there are organization changes.

We can visualize this stability using a decay curve, as seen in Figure 15. For rules at the user level, 56% exist for only a single week. Less than 1% of the rules exist for at least 14 weeks and only 0.07% of the rules exist for the entire study period. This high decay in how long rules last has several possible causes, including the possibility that the organization is changing rapid or simply that departments do not interact every week.

We know that VUMC is a large academic medical center, in which residents rotate through different clinical areas and in which care is not so much delivered by specific individuals but rather by care teams. Therefore, we can look at abstractions of the users to see if the delivery is by similar clinicians. For example, is it important that any anesthesiologist treats a patient, or must a specific anesthesiologist treat the patient? Since the only practical abstraction dataset we obtained is the department information, we use the abstraction of the user into his or her department.

When we abstract to the department level, we see a much slower rate of decay of the rules (Figure 15). Instead of the 44% drop after one week at the user level, the department level only had a 17% drop. While the department level still decays, the decay is much slower, such that 50% of the rules still remain at 7 weeks. Over 16% of rules exist the entire 21-week study period. This 16% represents a consistent and highly stable set of rules, which could be one group of rules to focus on when building an access control system or auditing system.

The highly dynamic nature of VUMC indicates that even at the abstracted level, any statistical model used for access control or auditing would need to be re-tuned or

regenerated on a fairly regular basis using the latest access data.

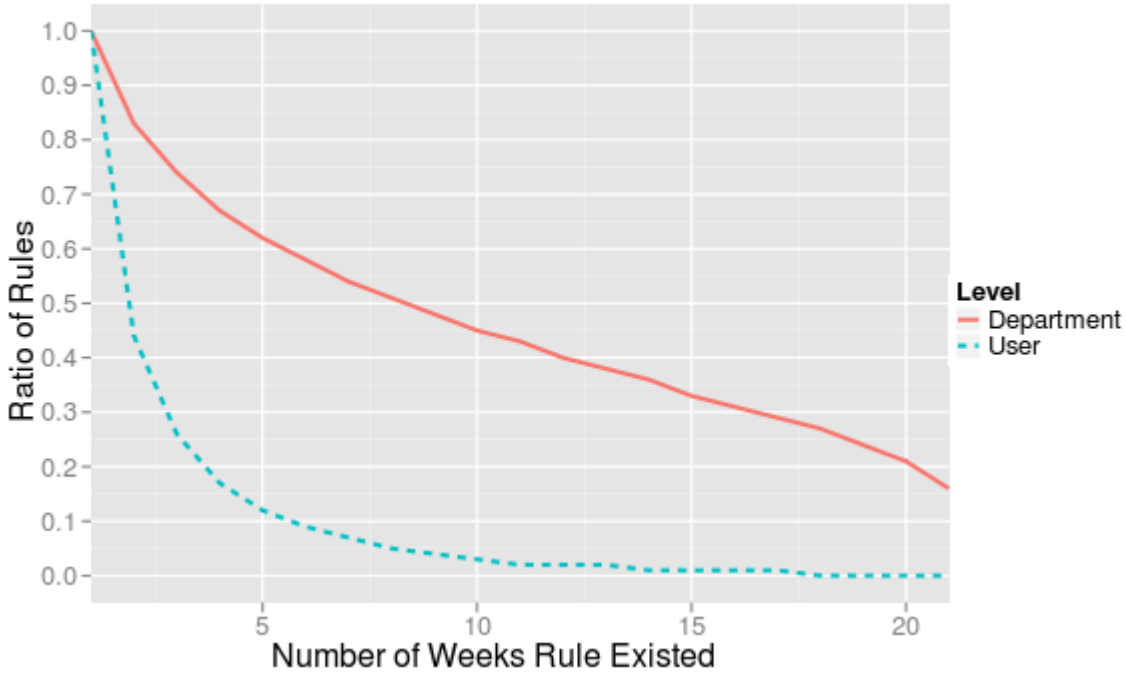

Figure 15: Decay of rules over time. 16% of the department rules existed for all 21-weeks of the study, while only 0.07% of the user rules lasted that long.

Sample Rules

We can examine a subset of these rules that HORNET generates from the StarPanel access logs to gain perspective on what the probabilities look like. We will look at just the department-level rules, since the user-level rules are de-identified and therefore lacking any context. Additionally, the department-level rules are much more stable, therefore more likely to be used by any anomaly detection system.

For our 21-week study period, we generated a total of 58,415 department-department rules in the form of pair and conditional probabilities of those departments accessing the records of the same patients.

High Conditional Probability Rules

The rules with the highest conditional probabilities, in which we have the probability of one department given some other department, as shown in Table 4, occurred for only a small part of the 21-week period (1 to 2 weeks). Most high conditional probability rules occur for a very short time—it is not until the 184th highest conditional probability rule that we see a rule that occurred for all 21 weeks of the study period. Therefore, we will use a simple filter to try to obtain rules with some permanence. If we only look at rules that occur for at least three weeks, we will eliminate any accesses that occurs in a single work session. Two weeks is too short since users can access the system across our arbitrary week-long periods, for example a shift that spans Saturday evening and Sunday morning would count as 2 weeks. An alternative would be to use a sliding window approach instead of our strict 7-day splits. Applying this exclusion criteria, we obtain Table 5, which has the highest conditional probability rules that occurred for at least 3 weeks of our 21-week study period. We see rules that are intuitive, such as the 0.378 probability of the Nephrology Clinic accessing a patient who was also accessed by the Hypertension Clinic. This result is intuitive since nephrology, the branch of internal medicine that focuses on the kidneys, often deals with kidney diseases that include hypertension, so we expect overlap between these two departments.

High Pair Probability Rules

While the rules with the highest conditional probabilities often do not appear for the full study period, the rules with the highest pair probabilities last for the entire study

period (Table 6). Since the pair probability represents the frequency a specific rule, we expect departments who traditionally see many patients at the top of the list of high pair probability rules. This expectation is validated by the departments in Table 6, such as the Emergency Medicine department. Furthermore, in Table 6 there are multiple rules where the antecedent is the same as the consequent. In this case, we are seeing the fact that people within the department are working very frequently with other people in that same department—a logical occurrence. These top 20 rules form a large proportion of the total number of rules in the 21-week study period, for example the "Emergency Medicine – Emergency Medicine" rule accounts for nearly 6.5% of all 58,415 rules detected in our study period. While a large number of the high pair probability rules are for the same department, some have different departments such as the rule between the Allergy/Pulm/Critical Care department and Emergency Medicine department, which represents 1.3% of all the rules.

Low Conditional Probability Rules

Just as we can see inherent organizational structure in the rules with the highest conditional and pair probabilities, we can also see this structure in the rules with the lowest conditional and pair probabilities. The rules with the lowest conditional probabilities (Table 7), applying the same 3-week minimum criterion, show department combinations that have low pair probabilities. The Emergency Medicine department is a common antecedent for many of these rules, since the Emergency Medicine department is connected to so many other departments (we are seeing the Emergency Medicine

department's role as a triage point into the hospital).

Low Pair Probability Rules

While the rules with the lowest conditional probabilities should likely be excluded when building a probabilistic model of VUMC, the rules with the lowest pair probabilities (Table 8) could possibly remain in the model. A low pair probability does not necessarily indicate that the rule is noise, but rather that it merely does not occur a large number of times in the total body of 58,415 rules.

Table 4: 20 Department Rules with Highest Conditional Probabilities. Includes the average and standard deviation of the conditional, P(B|A), and pair, P(A, B), probabilities over the 21 one-week periods. The "Weeks" column refers to how many of the one-week periods that rule appeared in. The department labels are from the clinician database.

		P(B|A)		P(A,B)		
A	**B**	**Average**	**Std Dev**	**Average**	**Std Dev**	**Week**
Nuclear Med Housestaff	Psychiatric Hosp at Vanderbilt	1	0	8.00E-06	0	1
VMG Bus Ofc Provider Enrollmnt	Pediatric Cardiology	1	0	5.06E-06	0	1
Ctr of Occupational & Env Med	Trauma	0.69	0.31	7.00E-06	9.96E-07	2
VCOEM	Trauma	0.69	0.31	7.00E-06	9.96E-07	2
Ctr for Molecular Neuroscience	Internal Medicine	0.5	0	1.90E-06	0	1
Psychology & Human Devel	Psychiatric Hosp at Vanderbilt	0.5	0	8.22E-06	0	1
Ctr for Molecular Neuroscience	Medicine	0.5	0	1.90E-06	0	1
VMG Bus Ofc Provider Enrollmnt	Pediatric Urology	0.5	0	2.64E-06	0	1
Nursing Education & Developmen	Rheumatology	0.5	0	1.92E-05	0	1
Computer Administration	Vanderbilt Home Care	0.5	0	1.60E-05	0	1
Nursing Education & Developmen	Rheumatology Clinic	0.5	0	1.92E-05	0	1
Human & Organizational Dev	School Of Nursing	0.4	0.23	3.60E-05	1.87E-05	7
Hypertension Clinic	Nephrology Clinic	0.38	0.17	4.16E-05	2.02E-05	3
VMG Bus Ofc Correspondence	The Learning Center	0.33	0.47	5.71E-06	3.72E-06	3
VMG Bus Ofc Fees Group	Strategy & Transformation	0.33	0	8.00E-06	0	1
VMG Bus Ofc Fees Group	Admissions Office	0.33	0	8.23E-06	0	1
VMG Bus Ofc Fees Group	Diabetes/Endocrinology	0.33	0	1.64E-05	0	1
VMG Bus Ofc Provider Enrollmnt	Liver Transplant	0.25	0.25	3.99E-06	1.35E-06	2
Ctr of Occupational & Env Med	Nephrology & Hypertension	0.25	0	8.33E-06	0	1
Radiology Administration	VMG - Franklin	0.25	0	5.01E-06	0	1

Table 5: 20 Department Rules with Highest Conditional Probability That Existed at Least 3 Weeks. Includes the average and standard deviation of the conditional, P(B|A), and pair, P(A, B), probabilities over the 21 one-week periods. The "Weeks" column refers to how many of the one-week periods that rule appeared in. The department labels are from the clinician database.

		P(B|A)		P(A,B)		
A	**B**	**Average**	**Std Dev**	**Average**	**Std Dev**	**Week**
Human & Organizational Dev	School Of Nursing	4.04E-01	2.28E-01	3.60E-05	1.87E-05	7
Hypertension Clinic	Nephrology Clinic	3.78E-01	1.73E-01	4.16E-05	2.02E-005	3
VMG Bus Ofc Correspondence	The Learning Center	3.35E-01	4.70E-01	5.71E-06	3.72E-06	3
VMG Bus Ofc Fees Group	Ofc of Compliance & Corp Integ	2.46E-01	1.84E-01	7.94E-06	2.21E-07	4
Main OR - Rollup	VMG - Franklin	1.94E-01	8.02E-02	5.94E-05	3.07E-05	6
Psychology & Human Devel	Mental Health Center	1.92E-01	7.55E-02	7.74E-06	4.77E-06	7
Human & Organizational Dev	Univ Comm Health Services	1.50E-01	7.84E-02	1.13E-05	4.94E-06	7
Main OR - Rollup	Spring Hill WIC	1.24E-01	5.90E-02	3.94E-05	2.12E-05	6
VMG Bus Ofc Fees Group	Internal Medicine	1.22E-01	1.50E-01	5.32E-06	2.06E-06	3
Radiology-Housestaff	Orthopaedics & Rehab	1.19E-01	1.00E-01	2.12E-05	1.23E-05	10
VMG Bus Ofc Fees Group	School Of Nursing	1.12E-01	7.83E-02	6.25E-06	2.58E-06	3
Radiology-Housestaff	Ob-Gyn	1.08E-01	6.60E-02	3.12E-05	1.95E-05	3
Pediatric Services Development	Neuro-Peds	1.06E-01	1.21E-01	4.52E-06	1.65E-06	3
Radiology Administration	Anesthesiology	1.06E-01	1.24E-01	6.19E-06	2.27E-06	9
VMG Bus Ofc Provider Enrollmnt	Neurosurgery	1.02E-01	1.66E-01	1.20E-05	9.12E-06	14
Psychology & Human Devel	Psychiatry	1.02E-01	7.69E-02	3.16E-06	1.38E-06	7
Psychology & Human Devel	Orthopaedics & Rehab	9.41E-02	1.10E-01	2.00E-06	8.48E-07	3
Psychology & Human Devel	Emergency Medicine	9.41E-02	1.10E-01	2.00E-06	8.48E-07	3
Institute of Imaging Science	Emergency Medicine	9.28E-02	1.33E-01	8.90E-05	5.53E-05	11
Junior League Unit	Pediatric Endocrinology	9.16E-02	1.02E-01	2.11E-04	1.28E-04	19

Table 6: 20 Department Rules with the Highest Pair Probabilities. Includes the average and standard deviation of the conditional, P(B|A), and pair, P(A, B), probabilities over the 21 one-week periods. The "Weeks" column refers to how many of the one-week periods that rule appeared in. The department labels are from the clinician database.

		P(B|A)		P(A,B)		
A	**B**	**Average**	**Std Dev**	**Average**	**Std Dev**	**Week**
Emergency Medicine	Emergency Medicine	6.47E-04	4.69E-05	6.48E-02	8.97E-03	21
Ophthalmology	Ophthalmology	3.36E-03	6.43E-04	2.85E-02	4.68E-03	21
Ob-Gyn	Ob-Gyn	1.55E-03	1.69E-04	2.69E-02	2.94E-03	21
Orthopaedics & Rehab	Orthopaedics & Rehab	1.47E-03	1.72E-04	2.06E-02	2.93E-03	21
Pediatric Hematology	Pediatric Hematology	4.28E-03	4.89E-04	2.06E-02	2.53E-03	21
Emergency Medicine	Emergency Med-Housestaff	1.75E-04	3.49E-05	1.73E-02	2.63E-03	21
Emergency Med-Housestaff	Emergency Medicine	2.56E-03	3.52E-04	1.73E-02	2.63E-03	21
School Of Nursing	School Of Nursing	1.83E-03	1.93E-04	1.70E-02	1.32E-03	21
Hematology/Oncology	Hematology/Oncology	2.01E-03	3.51E-04	1.59E-02	2.51E-03	21
Univ Comm Health Services	School Of Nursing	5.08E-03	1.07E-03	1.44E-02	1.72E-03	21
School Of Nursing	Univ Comm Health Services	1.56E-03	2.20E-04	1.44E-02	1.72E-03	21
Pediatric Cardiology	Pediatric Cardiology	2.71E-03	2.88E-04	1.43E-02	2.51E-03	21
Orthopaedics & Rehab	Emergency Medicine	1.02E-03	1.08E-04	1.42E-02	1.59E-03	21
Emergency Medicine	Orthopaedics & Rehab	1.44E-04	2.10E-05	1.42E-02	1.59E-03	21
School Of Nursing	Ob-Gyn	1.42E-03	1.89E-04	1.32E-02	1.60E-03	21
Ob-Gyn	School Of Nursing	7.63E-04	1.05E-04	1.32E-02	1.60E-03	21
Emergency Medicine	Pediatrics-Housestaff	1.32E-04	3.30E-05	1.30E-02	2.84E-03	21
Pediatrics-Housestaff	Emergency Medicine	2.49E-03	3.18E-04	1.30E-02	2.84E-03	21
Cardiovascular Medicine	Cardiovascular Medicine	1.58E-03	1.67E-04	1.30E-02	2.42E-03	21
Allergy/Pulm/Critical Care	Emergency Medicine	1.62E-03	1.98E-04	1.30E-02	1.55E-03	21

Table 7: 20 Department Rules with the Lowest Conditional Probabilities That Existed at Least 3 Weeks. Includes the average and standard deviation of the conditional, P(B|A), and pair, P(A, B), probabilities over the 21 one-week periods. The "Weeks" column refers to how many of the one-week periods that rule appeared in. The department labels are from the clinician database.

		P(B|A)		P(A,B)		
A	**B**	**Average**	**Std Dev**	**Average**	**Std Dev**	**Week**
Emergency Medicine	4 S GYN Holding/PACU	2.25E-08	6.69E-09	2.31E-06	7.38E-07	5
Emergency Medicine	Kennedy Center	2.29E-08	1.07E-08	2.00E-06	8.48E-07	3
Emergency Medicine	Psychology & Human Devel	2.29E-08	1.07E-08	2.00E-06	8.48E-07	3
Emergency Medicine	Cardiology 7N	2.35E-08	1.22E-08	2.40E-06	1.21E-06	9
Emergency Medicine	Pulmonary Clinic	2.93E-08	1.36E-08	2.73E-06	1.24E-06	4
Emergency Medicine	Admin MH Clinic	3.78E-08	2.13E-08	3.42E-06	2.05E-06	4
Emergency Medicine	Office Of Research	4.34E-08	5.77E-09	4.06E-06	3.24E-08	5
Emergency Medicine	NICU	5.09E-08	3.23E-08	5.00E-06	2.83E-06	9
Emergency Medicine	Worklife Connections -EAP	8.32E-08	1.01E-08	8.11E-06	7.63E-08	3
Emergency Medicine	Vanderbilt Oncology	8.54E-08	4.18E-08	8.07E-06	4.26E-06	4
Emergency Medicine	Radiology Administration	9.24E-08	6.30E-08	1.02E-05	8.25E-06	8
Emergency Medicine	Sedation Service	9.55E-08	6.67E-08	1.05E-05	8.78E-06	7
The Learning Center	Law School Deans Office	9.97E-08	2.63E-08	1.97E-06	7.44E-07	6
The Learning Center	University Library	9.97E-08	2.63E-08	1.97E-06	7.44E-07	6
Anesthesiology	Admin MH Clinic	1.03E-07	1.43E-08	1.58E-06	3.54E-08	5
Emergency Medicine	Urology/Pediatric Surgical Sci	1.10E-07	7.09E-08	9.56E-06	5.96E-06	4
Emergency Medicine	Special Procedures	1.12E-07	9.95E-08	1.02E-05	8.08E-06	8
Emergency Medicine	Ob/Gyn Practice	1.13E-07	7.08E-08	1.19E-05	7.98E-06	13
Emergency Medicine	Med Ethics	1.13E-07	7.95E-08	1.06E-05	6.43E-06	11
Emergency Medicine	VIPPS	1.14E-07	8.39E-08	1.12E-05	7.93E-06	17

Table 8: 20 Department Rules with the Lowest Pair Probabilities That Existed at Least 3 Weeks. Includes the average and standard deviation of the conditional, P(B|A), and pair, P(A, B), probabilities over the 21 one-week periods. The "Weeks" column refers to how many of the one-week periods that rule appeared in. The department labels are from the clinician database.

		P(B|A)		P(A,B)		
A	**B**	**Average**	**Std Dev**	**Average**	**Std Dev**	**Week**
Nursing Support Services	7-S MICU	1.18E-05	3.61E-06	5.11E-07	1.78E-07	3
7-S MICU	Nursing Support Services	1.37E-06	6.90E-07	5.11E-07	1.78E-07	3
9S General Surgery	Special Procedures	4.45E-06	2.17E-06	5.81E-07	4.52E-09	3
9S General Surgery	Hematology/Stem Cell Clinic	4.45E-06	2.17E-06	5.81E-07	4.52E-09	3
Special Procedures	9S General Surgery	4.98E-04	4.78E-04	5.81E-07	4.52E-09	3
Hematology/Stem Cell Clinic	9S General Surgery	1.67E-06	1.64E-07	5.81E-07	4.52E-09	3
9S General Surgery	Radiology Administration	4.23E-06	1.91E-06	5.82E-07	4.91E-09	4
Radiology Administration	9S General Surgery	3.56E-04	4.66E-04	5.82E-07	4.91E-09	4
Nursing Support Services	SICU	1.42E-05	3.39E-06	6.48E-07	3.72E-07	3
SICU	Nursing Support Services	8.07E-06	7.40E-06	6.48E-07	3.72E-07	3
Asthma/Sinus/Allergy Prog	SICU	7.94E-05	2.60E-05	6.74E-07	1.07E-08	4
SICU	Asthma/Sinus/Allergy Prog	5.00E-06	1.30E-06	6.74E-07	1.07E-08	4
Special Procedures	Nursing Support Services	4.46E-04	4.24E-04	6.86E-07	1.82E-07	4
Hematology/Stem Cell Clinic	Nursing Support Services	2.08E-06	7.39E-07	6.86E-07	1.82E-07	4
Nursing Support Services	Special Procedures	1.12E-05	2.75E-06	6.86E-07	1.82E-07	4
Myelosuppresion (11N)	Nursing Support Services	9.48E-06	1.72E-06	6.86E-07	1.82E-07	4
Nursing Support Services	Myelosuppresion (11N)	1.12E-05	2.75E-06	6.86E-07	1.82E-07	4
Nursing Support Services	Hematology/Stem Cell Clinic	1.12E-05	2.75E-06	6.86E-07	1.82E-07	4
9S General Surgery	SICU	3.72E-06	2.08E-06	7.03E-07	3.36E-07	4
SICU	9S General Surgery	8.48E-06	6.45E-06	7.03E-07	3.36E-07	4

Discussion

In this chapter we demonstrated our use of HORNET to generate a statistical model of how VUMC operates from the access logs of its electronic medical record system. This model includes a wide range of access patterns—from heavy users to occasional users, from patients who are rarely accessed to patients who are commonly accessed. We also provided some justification for who these users are in relation to their access patterns. Additionally, our results show that VUMC is a dynamic organization, but we see greater stability when we model the users in terms of their department affiliations.

This model and the specific results from our 21-week study period are not the major finding in this chapter. The major finding in this chapter is that our tool, HORNET, can effectively and efficiently generate a probabilistic model of a Healthcare Organization from the access logs of that organization's information system. This tool seamlessly incorporates any meta-information about the organization, such as human-resource type data, even if that data is less than ideal. Furthermore, it is important to realize that the specific results in this chapter, such as the rules, are not meant to be applicable at any other organization or even at VUMC today, three and a half years after this data set was collected—especially given the rate of decay of rules. This lack of current applicability or applicability to other organizations does not mean the results are not reproducible. Far from it, with our open source HORNET tool, any organization is able to generate a customized statistical model of itself.

Limitations

While we successfully built a model from StarPanel's access logs, we do have several limitations—some related to our approach and some related to the realities of working with production data in a way that was never intended.

From our perspective, the biggest limitation is from our co-opting of production data and use of it in ways for which is was never designed. Having complete, consistent, accurate, and clean meta-information on all users in the system would be very beneficial. In addition to using the department meta-information, we could also use roles. We believe that the combination of role and department would have made a very compelling statistical model (for example, do attending physicians in one department always work with charge nurses in some other department?). Alas, we are unable to find the "perfect" source of meta-information. Due to the nature of what various administrators were trying to model, we know of at least 4 distinct sources of partial meta-information. The first source is a database collected to keep track of the medical residents. Since the residents often rotate through different care areas in the organization, this database does not contain information of what department or care area each resident is in or what time they were in that area. The second source is designed to allow customization of experience within StarPanel based on a self-selected role—unfortunately, a majority of users never customize their experience. A third source, while containing department information, lacks any key to which we could link the users in this source to the users in the access logs. Additionally, this third source holds different department labels than the primary meta-information source we use. We do not wish to perform a mapping between the two

sets of labels (while linking "Emergency Medicine" to "Emergency Med" is straightforward, there are cases in which we lack the domain knowledge to make an informed match). We doubt this is an issue that is specific to VUMC—the effort to create and maintain a complete meta-information source often outweighs the potential benefits of such a database.

Interestingly, this is an issue that any role-based security system in healthcare must itself confront. Due to the lack of complete, centralized knowledge about who users are, any role-based security system must first undertake the monumental task of creating and maintaining such a database. While a probabilistic anomaly detection system based upon the work presented in this document would ideally have such a complete source of meta-information, a detection system using HORNET could be greedy and take whatever various sources are available. For example, we could generate one model based upon the resident database and apply it to detect anomalies for only the residents, then use a more standard employee database to detect anomalies for only those employees. This greedy approach would allow HCO administrators to more quickly and cheaply deploy an anomaly detection system, since the complete meta-information database would never have to be created.

For this study, we ignore a significant portion of the accesses when abstracting the model to the department level, since we lack department information for a large number of our users. While less than ideal, we believe that the exclusion does not negatively affect the model. We believe so because the users with departments account for a larger portion of the accesses and the two groups (those with and without departments) are, as

we showed, indistinguishable in their access patterns.

While the lack of complete meta-information is the biggest limitation of this study, we have several limitations specific to our modeling technique. For one, we only build the relational networks for one week periods, instead of trying to vary the period size or define periods as the length of individual stays in the hospital based on admission and discharge. We noticed a strong 7-day periodicity in the data (Figure 10), and chose a logical default value.

Another limitation is that we perform only a single type of weighting on the relationships, when there are several additional logical ways to weight the relationships. We do weight the relationships based upon how many other relationships a user is involved in, such that the relationship for a user who has only a single relationship is stronger that a relationship for a user who has relationships with 100 other users. In addition to this style of weighting, we can discount patients who are accessed by a large number of users, using a weighting system similar to term frequency-inverse document frequency that is commonly found in information retrieval.

We could also weight the rules by strength (either conditional or pair probability), allowing us to understand if important rules disappear or if the important rules last all 21 weeks (Figure 15).

CHAPTER VI

SURVEY OF EXPERT UNDERSTANDING OF ORGANIZATION

In the previous chapter, we evaluated a simple, yet powerful, method for obtaining a statistical model how clinicians interact via patients' medical records. This model represents actual clinician usage of the medical record system and can be the basis for automatically defining a security policy to restrict access to medical records or detect possible improper accesses. Traditional methods for defining the security policies rely upon organizational experts manually defining these policies. Since our model offers a potential replacement or supplement to these traditional methods, we wish to understand how well experts perform at characterizing how the organization delivers care. To compare the HORNET data mining approach, we administered a pilot, which demonstrates the benefits of automatically mining access control and lays the groundwork for a more thorough study on the subject.

Survey Methods

Survey Design

This preliminary survey is designed to test the null hypothesis that experts are good at approximating the conditional probability of two departments caring for the same patients, a task necessary for manually building access control policies. To test this

hypothesis, we randomly select 3 departments from the set of rules presented in the previous chapter: the SICU, Ob-Gyn, and the Nephrology & Hypertension departments. For each of these 3 departments, we select 10 rules in which the department is the antecedent in the conditional probability, therefore creating a total of 30 rules (as shown in Table 9, along with the conditional and pair probabilities). Each set of 10 rules is from a cross section of the conditional probability distribution of rules involving the antecedent, such that we pick roughly one rule from every 10^{th} percentile (e.g. ~99^{th} percentile, ~90^{th} percentile, ~80^{th} percentile, etc.). Additionally, all the rules occur for at least 16 weeks in our 22 week study period. We arbitrarily set the 16 week limit to ensure that all the rules in the survey consistently occur within the organization. For each of the 30 rules we attach a 1-to-5 Likert-scale question, for example:

"**If a patient is seen by the 'SICU' department, how likely is that patient to be seen by the '5N CVICU' department?"**

In an introduction to the survey (Appendix A), we ask the survey participants to rate each statement on a 1 to 5 scale based upon their own knowledge of the organization, where 5 indicates they think it is very likely and 1 indicates they think it is unlikely. If experts are good at defining the conditional probabilities, we would expect to see high scores for rules with high conditional probabilities, and low scores for rules with low conditional probabilities.

Administration

Using the RedCap Survey system we included an introduction, the 30 rules and

their Likert scale choices, and an optional comments box, in a web-based survey, using the RedCap Survey tool (52). We solicited responses from 15 attending physicians at VUMC. These physicians have appointments in a variety of departments including Internal Medicine, Emergency Medicine, and various Intensive Care Units. We used the RedCap system to email the survey to the individuals and anonymously collect responses. We received 7 responses over the course of one month, resulting in a 44% response rate.

Analysis

To test our hypothesis, we try to find a linear correlation between the conditional probabilities of the 30 rules against the average score of the respondents. If a correlation exists, we can assume that the experts agreed with the probabilistic model that HORNET generated. If no correlation or a very poor correlation exists, we question the ability of human experts to document how the organization works. Additionally, we assess for inter-rater agreement using Fleiss' kappa (53) and look at the deviation for each survey question. We use Fleiss' kappa instead of the more traditional Cohen's kappa since the survey has multiple choices for each question and we wish to compare all respondents. One weakness of using Fleiss' kappa is that we treat the Likert scale as a categorical variable instead of an ordinal value. Thus we lose the context that a response of 4 is more similar to 3 or 5 than it is to 1 or 2. Therefore, to supplement Fleiss' kappa, we examine the deviation of the respondents' compared to the mean.

It is important to note that results of the survey are not sufficiently powered to make any formal claims. Yet, as a pilot study, it can provide suggestions of possible

results and directions for future, more powerful studies.

Results

We can see the descriptive statistics of how the 7 respondents answered each question in Table 9 along with the probabilities that HORNET computed. From this data, we can compute the inter-rater agreement. For all three antecedent groups (SICU, Ob-Gyn, and Nephrology & Hypertension), Fleiss' kappa showed there was only slight agreement between the raters (54), since all three had kappa values in the range of 0.11 to 0.18. The SICU group had a kappa of 0.11, the Ob-Gyn group had a kappa of 0.18, and the Nephrology & Hypertension group had a kappa of 0.16. However, by using Fleiss' kappa we are treating the responses as categorical data instead of ordinal data, so we lose the context of the Likert scale. If we examine the raw standard deviations we see that 20 of the 30 response standard deviations are less than 1 step in the Likert scale. Examining the standard deviations suggests that, while not strong, there is at least a degree of agreement between the raters.

Since we wish to understand if there is a correlation between the Percentile of P(B|A) and the Survey Response for each department pair, we try to fit a linear model to the data. Plotting the Percentile of the conditional probability of the rules against the Survey Response (Figure 16), yields a very poor result. We expect that for increasing conditional probabilities (represented by the increasing percentile) we should see an increase in the expert's belief in the likelihood of the relationship between the departments. However, as Figure 16 shows, only the Nephrology & Hypertension has a

successful linear fit, even though it is a poor fit with a R^2 value of 0.38.

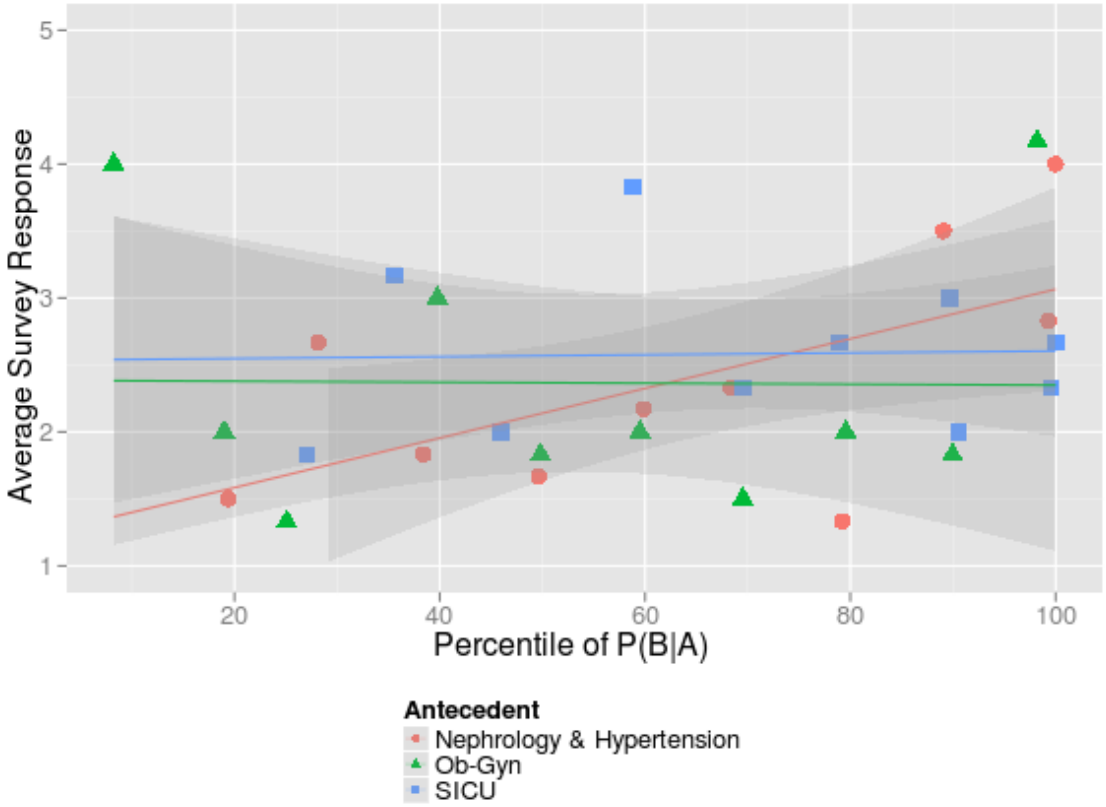

Figure 16: Survey Responses versus Conditional Percentile.

Table 9: Survey Results

		Probabilistic Model				Survey Response			
A	B	P(B|A)	P(A,B)	Weeks	Percentile of P(B|A)	Mean	Std Dev	Coefficient of Variance	Median
SICU	Trauma PCC	1.90E-03	2.04E-04	20	100	2.86	1.07	0.71	3
SICU	7-S MICU	3.14E-04	3.36E-05	21	90.56	2	0.58	1	2
SICU	Allergy/Pulmonary/Critical Care	1.84E-03	1.86E-04	21	99.57	2.29	0.76	1.14	2
SICU	Transplant Center	2.76E-04	3.14E-05	21	89.7	3.14	1.35	0.79	3
SICU	Diabetes/Endocrinology	1.19E-04	1.34E-05	20	78.97	2.86	1.21	0.71	3
SICU	Neurology	8.16E-05	8.65E-06	21	69.53	2.43	0.79	0.81	2
SICU	TVC PreAdmit Testing	5.41E-05	5.59E-06	19	58.8	4	0.82	0.8	4
SICU	Psychiatry	3.59E-05	4.07E-06	18	45.92	2.14	0.69	0.71	2
SICU	5N CVICU	2.94E-05	3.05E-06	17	35.62	3.29	0.76	0.82	3
SICU	Neuro-Sleep Disorders	2.13E-05	2.51E-06	17	27.04	1.86	0.38	0.93	2
Ob-Gyn	Anesthesiology	1.70E-04	2.97E-03	21	98.21	4.29	0.76	0.86	4
Ob-Gyn	Pediatric Cardiology	4.59E-05	7.93E-04	21	89.96	1.71	0.95	1.71	1
Ob-Gyn	Plastic Surgery	1.76E-05	3.08E-04	21	79.57	2.14	0.69	0.71	2
Ob-Gyn	Pediatric Critical Care	8.77E-06	1.53E-04	21	69.53	1.43	0.79	1.43	1
Ob-Gyn	Pediatric Urology	3.74E-06	6.68E-05	20	59.5	1.86	1.21	1.86	1
Ob-Gyn	Trauma PCC	2.36E-06	4.23E-05	21	49.82	2	0.58	0.67	2
Ob-Gyn	Pain Center	1.63E-06	2.72E-05	20	39.78	3	0.58	1	3
Ob-Gyn	5S PICU	9.38E-07	1.62E-05	20	25.09	1.29	0.49	1.29	1
Ob-Gyn	Asthma/Sinus/Allergy Program	6.96E-07	1.25E-05	17	19	2.29	0.95	0.57	2
Ob-Gyn	Nursing Support Services	3.72E-07	6.42E-06	18	8.24	4	1	1	4

Table 9, continued

		Probabilistic Model				Survey Response			
A	B	P(B|A)	P(A,B)	Weeks	Percentile of P(B|A)	Mean	Std Dev	Coefficient of Variance	Median
Nephrology & Hypertension	Emergency Medicine	1.65E-03	1.16E-02	21	100	4.14	1.07	0.83	4
Nephrology & Hypertension	Allergy/Pulmonary/Critical Care	4.81E-04	3.38E-03	21	99.3	3	1.15	0.75	3
Nephrology & Hypertension	Renal Transplant	9.37E-05	6.40E-04	21	89.08	3.57	1.27	0.89	4
Nephrology & Hypertension	Neonatology	3.84E-05	2.61E-04	21	79.23	1.29	0.49	1.29	1
Nephrology & Hypertension	Oncology/Hematology	2.02E-05	1.39E-04	21	68.31	2.43	0.79	0.81	3
Nephrology & Hypertension	Psychiatry	1.12E-05	7.89E-05	21	59.86	2.43	0.98	0.61	2
Nephrology & Hypertension	Oral Surgery	6.31E-06	4.57E-05	20	49.65	2	1	0.5	2
Nephrology & Hypertension	Pediatric Emergency	3.92E-06	2.63E-05	19	38.38	1.71	1.25	1.71	1
Nephrology & Hypertension	Cardiac Cath Lab	2.37E-06	1.61E-05	16	28.17	2.86	0.9	0.71	3
Nephrology & Hypertension	Student Health & Wellness	1.57E-06	1.14E-05	16	19.37	1.43	0.79	1.43	1

Discussion

From these results, we can draw two preliminary conclusions: first, human domain experts perform poorly at manually building a probabilistic model of how the medical center operates, and second, the agreement between experts is less than ideal, thus raising questions of how one would ever correctly pick experts if manually building such a model. We believe that the poor expert performance indicates the superiority of a computational approach to generate this type of organizational model of the organization. However, we believe that the use of our approach in addition to using experts in tasks they are more suited for, such as defining policies like "a user should not access a coworker's record," will perform better than our approach alone. As stated previously, the survey was not fully powered, thus we believe that further, more thorough studies should be conducted to validate this preliminary results.

While the results are interesting in that they suggest the need to use a tool such as HORNET to build a probabilistic model of the organization as compared to the traditional, human-based method of building intrusion detection rules or access control policies, they do have two possible flaws.

The first possible flaw is in the survey instrument. While we performed several iterations of survey development and testing, multiple respondents indicated in the free response section that it was difficult to think through what the survey was asking for. We attempted to model the survey on the task that an administrator would have to conduct in order to build policies for access control or intrusion detection. Improvements can likely be made to clarify the survey device, but we believe that the respondents' difficulty in

completing the survey indicates the inherent difficulty of using humans to quantitatively characterize how a complex organization works.

The second flaw is related to the issue first discussed in the previous chapter about the secondary use of meta-information data sources. Several respondents indicated that they were unsure of what the department labels referred to. It is likely that at least some of the department labels serve an administrative purpose that most clinicians would be unaware of. For example "Nursing Support Services" refers to an administrative department label, but would any practicing clinician know who is part of this department? This flaw helps to further advance our previous arguments that complete and accurate meta-information databases should be built to provide this information universally and that traditional methods would run into the same issue. This flaw also demonstrates the superiority of computational approaches to building the model using a bottom up approach instead of what a manual, human based system would perform.

CHAPTER VII

DISCUSSION & FUTURE WORK

In the previous chapters, we introduced an open-source tool for building access logs into a relational network, from which we can obtain a statistical model of how a healthcare organization operates. Using this tool, we successfully demonstrated its use on access logs from StarPanel, Vanderbilt University Medical Center's EMRS. We additionally measured the performance of domain experts in building a similar statistical model.

This work comes at an opportune time in a field that has a major unsolved problem. To date, healthcare organizations have been largely unsuccessful with implementing solutions that protect patient privacy against insider attacks. Neither auditing nor access control has been truly deployed in HCOs, due to numerous issues of defining the models of what actions are and are not allowed. These issues include human inability to characterize the models and natural, constant structural changes of the organization. In spite of these issues, HCOs now have a mandate from the government to increase the adoption and security of electronic healthcare systems. Our work provides a potential first step to a solution of these issues by automatically learning how the organization naturally operates.

Specifically, we have several contributions to the field. First, and most importantly, is our Apache-licensed tool itself. Administrators at other HCOs can now

conduct the same analysis using data from their own institution. HORNET is also a platform that researchers and software developers can build on top of, taking advantage of the system's advanced plugin architecture, configuration system, and plugin chaining.

Second, we contribute an approach which makes the best of incomplete data. Our findings suggest that we can successfully find meaningful results in spite of missing data. Existing methods of policy definitions would almost certainly need complete, accurate, and maintained data in order to function. Our approach will accept whatever information is available about the organization. However, we argue that for the best performance of any access control system or auditing system, administrators should work to assemble and maintain a complete, detailed, and accurate representation of the people who have access to health information systems.

Third, we provide pilot results of a survey that indicates the difficultly of using humans to manually define a statistical model for access control or auditing. This is not to say that humans are not good at defining certain classes of rules, for example, that Ob/Gyn physicians should not look at male patients' medical records. An ideal access control system would likely incorporate both humans and computer generated models—leaving each to what it is best at.

While our work focuses on healthcare, specifically on electronic medical record systems, there is a wider applicability for this work. Any system which logs a user accessing some resource could be modeled with our tool. In the example at Vanderbilt, patients were this resource. A natural extension is examining Clinical Provider Order Entry (CPOE) systems' access logs or other types of HIS. Even beyond the healthcare

realm, this approach and tool can be used to model use of web sites.

Next Steps

With the open-source distribution of our framework, the future of this work is open to the wishes of HCO administrators and researchers. Specifically, we plan to continue to develop HORNET, by improving its core, adding plugins for different types of analysis, and developing a graphical user interface.

In each chapter, where we have indicated limitations of our existing methods, we can work to fix these inadequacies. We plan to experiment with period sizes other than 7 days and with different methods of weighting relationships. Additionally, we can use the temporal context of the access logs to generate probabilities that have a temporal meaning. For example, we might expect that the Emergency Department sees patients before any other department.

If VUMC provides a complete source of meta-information about the users of StarPanel, including the users' department and role, we can make more specific statistical models.

We suspect that our model can be dramatically improved using several additional techniques and sources of information. Using referrals within the organization we could confirm existing work patterns, which might confirm and allow tuning of our statistical model. Additional improvements will likely result from incorporating a temporal view of the data (for example, we might expect the emergency medicine department to access records first, then some other department such as internal medicine or surgery). Our

model also only models pairwise relationships—the expansion of this into group detection using clustering techniques could provide more specific rules.

The most important next step in this work is incorporating the statistical model we dynamically generate into an access control system or into a retrospective auditing system. We specifically plan to incorporate our system into Vanderbilt's Model-Integrated Clinical Information System (MICIS) to define access control policies. Additionally, we plan to spike the access logs from our study period with known improper accesses to perform a retrospective audit.

APPENDIX A

EXPERT SURVEY

Expert Survey of Departmental Collabrations in Delivering Care (For Attendings)

Returning?

Dear Clinician,

Protecting a patient's privacy is a significant concern to all of us. I am a master's student in the Department of Biomedical Informatics and as part of my master's I developed a computer program that characterizes the relationships between StarPanel users, based upon which patients' StarPanel records they access in common.

As part of this research, I am surveying a small group of attending physicians and nurses at Vanderbilt University Medical Center (VUMC) to compare the computer program's understanding of departmental relationships at VUMC with your opinion about the strength or probability of those relationships. Our hope for the outcome of this research is to help VUMC protect patients' privacy. This survey will take you less than 10 minutes to complete.

Below are 30 statements, involving 3 departments. Based on your knowledge, for statement, please use the 5-point scale to indicate your impression of how likely that statement is, where 1 indicates that you expect to never see it, and 5 indicates that you strongly expect to see it.

For example,
How likely would a clinician in the 'Pediatric Cardiology' department care for a patient who has already or will be cared for by a clinician in the 'Pediatric Pulmonary'? 0 1 2 4 **5**

You would pick **5** because you believe that it is very likely that these two departments will care for the same patients.

If you have questions or comments, please contact me.

Thank you for your time and contribution!

John Paulett
john.paulett@vanderbilt.edu
615-936-2814

Department of Biomedical Informatics
Masters Committee:
Dr. Brad Malin
Dr. Nancy Lorenzi
Dr. Dario Giuse

1) **If a patient is seen by the 'SICU' department, how likely is that patient to be seen by the '5N CVICU' department?**

○ 1 ○ 2 ○ 3 ○ 4 ○ 5

REFERENCES

1. Kohn LT, Corrigan J, Donaldson MS. To Err Is Human: Building a Safer Health System. National Academy Press; 2000.

2. Corrigan JM. Crossing the Quality Chasm: A New Health System for the 21st Century. Washington, DC, National Academy Press; 2001.

3. Barrows RC, Clayton PD. Privacy, confidentiality, and electronic medical records. Journal of the American Medical Informatics Association. 1996 ;3(2):139-148.

4. Clayton PD, Boebert WE, Defriese GH, Dowell SP, Fennell ML, Frawley KA. For the Record: Protecting Electronic Health Information. National Academy Press; 1997.

5. Bernard HR, Killworth P, Kronenfeld D, Sailer L. The Problem of Informant Accuracy: The Validity of Retrospective Data. Annual Reviews in Anthropology. 1984 ;13(1):495-517.

6. Gallagher RJ, Sengupta S, Hripcsak G, Barrows RC, Clayton PD. An audit server for monitoring usage of clinical information systems. Proc AMIA Symp. 1998 ;1002

7. Shaw E, Ruby KG, Post JM. The Insider Threat to Information Systems. Security Awareness Bulletin. 1998 ;227-46.

8. Collingsworth B, Menezes R. Identification of Social Tension in Organizational Networks. In: Complex Networks: Results of the 1st International Workshop on Complex Networks (CompleNet 2009). Springer; 2009. p. 209.

9. Malin BA. Correlating Web Usage of Health Information with Patient Medical Data. Proc AMIA Symp. 2002 ;484-488.

10. Mathe JL, Duncavage S, Werner J, Malin B, Ledeczi A, Sztipanovits J. Implementing a Model-Based Design Environment for Clinical Information Systems [Internet]. In: Workshop on Model-Based Trustworthy Health Information Systems In Conjunction with Models 2007. 2007. Available from: http://www.truststc.org/pubs/308.html

11. Mathe JL, Werner J, Lee Y, Malin B, Ledeczi A. Model-based design of clinical information systems. [Internet]. Methods of Information in Medicine. 2008 ;Available from: http://www.truststc.org/pubs/377.html

12. Barth A, Datta A, Mitchell JC, Nissenbaum H. Privacy and Contextual Integrity: Framework and Applications. In: Proceedings of the 2006 IEEE Symposium on

Security and Privacy. 2006. p. 184-198.

13. Wang SJ, Middleton B, Prosser LA, Bardon CG, Spurr CD, Carchidi PJ, et al. A cost-benefit analysis of electronic medical records in primary care. The American Journal of Medicine. 2003 ;114(5):397-403.

14. Hillestad R, Bigelow J, Bower A, Girosi F, Meili R, Scoville R, et al. Can Electronic Medical Record Systems Transform Health Care?: Potential Health Benefits, Savings, And Costs. Health Affairs. 2005 ;24(5):1103.

15. Clooney hospital punishes staff [Internet]. BBC. 2007 Oct 10;[cited 2009 Jun 29] Available from: http://news.bbc.co.uk/2/hi/7037919.stm

16. Ornstein, Charles. UCLA workers snooped in Spears' medical records - Los Angeles Times [Internet]. [cited 2009 Jun 29] Available from: http://articles.latimes.com/2008/mar/15/local/me-britney15

17. Ornstein, Charles. UCLA staffer looked through Farrah Fawcett's medical records - Los Angeles Times [Internet]. [cited 2009 Jun 29] Available from: http://articles.latimes.com/2008/apr/03/local/me-farrah3

18. Konrad W. Medical Problems Could Include Identity Theft [Internet]. The New York Times. 2009 Jun 13;[cited 2009 Jun 28] Available from: http://www.nytimes.com/2009/06/13/health/13patient.html?_r=1&hpw

19. Johnson ME. Data Hemorrhages in the Health-Care Sector.

20. Denley I, Smith SW. Privacy in clinical information systems in secondary care. BMJ: British Medical Journal. 1999 ;318(7194):1328-1331.

21. Randazzo MR, Keeney M, Kowalski E, Cappelli D, Moore A. Insider Threat Study: Illicit Cyber Activity in the Banking and Finance Sector. DTIC Research Report ADA441249; 2005.

22. Office for Civil Rights - HIPAA [Internet]. Available from: http://www.hhs.gov/ocr/hipaa/

23. Collins, Eve. Calif. Hands Down Maximum $250,000 Penalty for Employee Snooping Into Patient Records at Kaiser Permanente Bellflower Medical Center [Internet]. 2009 Jun 10;[cited 2009 Jun 28] Available from: http://www.aishealth.com/Bnow/hbd061009.html

24. Steinbrook R. Health Care and the American Recovery and Reinvestment Act. The New England Journal of Medicine. 2009 ;360(11):1057.

25. Cooley R, Mobasher B, Srivastava J. Web Mining: Information and Pattern Discovery on the World Wide Web. Proceedings of the 9th IEEE International Conference on Tools with Artificial Intelligence (ICTAI'97). 1997 ;1(2.1):

26. Srivastava J, Cooley R, Deshpande M, Tan PN. Web usage mining: discovery and applications of usage patterns from Web data. ACM SIGKDD Explorations Newsletter. 2000 ;1(2):12-23.

27. Cooley R, Mobasher B, Srivastava J, others. Data Preparation for Mining World Wide Web Browsing Patterns. Knowledge and Information Systems. 1999 ;1(1):5-32.

28. Giles, Jim. Email patterns can predict impending doom - tech - 22 June 2009 - New Scientist [Internet]. 2009 Jun 22;[cited 2009 Jun 28] Available from: http://www.newscientist.com/article/mg20227135.900-email-patterns-can-predict-impending-doom.html

29. Nieder GL, Nagy F. Analysis of medical students' use of Web-based resources for a gross anatomy and embryology course. Clinical Anatomy. 2002 ;15(6):409-418.

30. D'Alessandro MP, D'Alessandro DM, Galvin JR, Erkonen WE. Evaluating overall usage of a digital health sciences library. Bull Med Libr Assoc. 1998 ;86(4):602-9.

31. Dev P, Rindfleisch TC, Kush SJ, Stringer JR. An analysis of technology usage for streaming digital video in support of a preclinical curriculum. Proc AMIA Symp. 2000 ;180180-184.

32. Chen ES, Cimino JJ. Automated discovery of patient-specific clinician information needs using clinical information system log files. AMIA Annu Symp Proc. 2003 ;145-9.

33. Malhotra S, Jordan D, Shortliffe E, Patel VL. Workflow modeling in critical care: Piecing together your own puzzle. Journal of Biomedical Informatics. 2007 ; 40(2):81-92.

34. Carley KM, Lee JS. Dynamic Organizations: Organizational Adaptation in a Changing Environment. Advances in Strategic Management. 1998 ;15269-297.

35. Maruster L, Aalst WVD, Weijters T, Bosch AVD, Daelemans W. Automated Discovery of Workflow Models from Hospital Data. Proceedings of the ECAI Workshop on Knowledge Discovery from Temporal and Spatial Data. 2002 ;32-37.

36. van der Aalst WMP, van Dongen BF, Herbst J, Maruster L, Schimm G, Weijters A. Workflow mining: A survey of issues and approaches. Data and Knowledge

Engineering. 2003 ;47(2):237-267.

37. Agrawal R, Imielinski T, Swami A. Mining association rules between sets of items in large databases. ACM SIGMOD Record. 1993 ;22(2):207-216.

38. Apache License, Version 2.0 - The Apache Software Foundation [Internet]. [cited 2009 May 26] Available from: http://www.apache.org/licenses/LICENSE-2.0.html

39. Python Programming Language -- Official Website [Internet]. [cited 2009 May 26] Available from: http://www.python.org/

40. Wasserman S, Faust K. Social Network Analysis: Methods and Applications. Cambridge University Press; 1994.

41. Travers J, Milgram S. An Experimental Study of the Small World Problem. Sociometry. 1969 ;32(4):425-443.

42. Milgram S. The small world problem. Psychology Today. 1967 ;2(1):60-67.

43. Leskovec J, Horvitz E. Worldwide Buzz: Planetary-Scale Views on an Instant-Messaging Network. In: Proc. 17th International World Wide Web Conference. 2008.

44. Bhagat S, Rozenbaum I, Cormode G. Applying link-based classification to label blogs. In: Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis. 2007. p. 92-101.

45. Malin B, Carley K. A Longitudinal Social Network Analysis of the Editorial Boards of Medical Informatics and Bioinformatics Journals. Journal of the American Medical Informatics Association. 2007 ;14(3):340-348.

46. Merrill J, Hripcsak G. Using Social Network Analysis within a Department of Biomedical Informatics to Induce a Discussion of Academic Communities of Practice. Journal of the American Medical Informatics Association. 2008 ;15(6):780-782.

47. Baeza-Yates R, Ribeiro-Neto B. Modern information retrieval. Addison-Wesley Harlow, England; 1999.

48. Hoot N, Weiss J, Giuse D, Jirjis J, Peterson J, Lorenzi N, et al. Integrating communication tools into an electronic health record. Medinfo. 2004 ;2004

49. Giuse DA. Supporting communication in an integrated patient record system. AMIA Annu Symp Proc. 2003 ;1065

50. Denny JC, Giuse DA, Jirjis JN. The Vanderbilt Experience with Electronic Health Records. In: Seminars in Colon and Rectal Surgery. Elsevier; 2005. p. 59-68.

51. Giuse DA, Mickish A. Increasing the availability of the computerized patient record. Proc AMIA Annu Fall Symp. 1996 ;633–637.

52. Harris PA, Taylor R, Thielke R, Payne J, Gonzalez N, Conde JG. Research electronic data capture (REDCap)—A metadata-driven methodology and workflow process for providing translational research informatics support. Journal of Biomedical Informatics. 2009 ;42(2):377-381.

53. Fleiss JL. Measuring nominal scale agreement among many raters. Psychological Bulletin. 1971 ;76(5):378-382.

54. Landis J, Koch G. The Measurement of Observer Agreement for Categorical Data. Biometrics. 1977 Mar ;33(1):174, 159.