Journal of Biomedical Informatics 44 (2011) 333–342

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

Learning relational policies from electronic health record access logs

Bradley Malin *, Steve Nyemba, John Paulett

Department of Biomedical Informatics, School of Medicine, Vanderbilt University, Nashville, TN, USA

ARTICLE INFO

Article history:
Received 3 August 2010
Available online 26 January 2011

Keywords:
Electronic health records
Organizational behavior
Knowledge discovery
Access logs
Auditing

ABSTRACT

Modern healthcare organizations (HCOs) are composed of complex dynamic teams to ensure clinical operations are executed in a quick and competent manner. At the same time, the fluid nature of such environments hinders administrators' efforts to define access control policies that appropriately balance patient privacy and healthcare functions. Manual efforts to define these policies are labor-intensive and error-prone, often resulting in systems that endow certain care providers with overly broad access to patients' medical records while restricting other providers from legitimate and timely use. In this work, we propose an alternative method to generate these policies by automatically mining usage patterns from electronic health record (EHR) systems. EHR systems are increasingly being integrated into clinical environments and our approach is designed to be generalizable across HCOs, thus assisting in the design and evaluation of local access control policies. Our technique, which is grounded in data mining and social network analysis theory, extracts a statistical model of the organization from the access logs of its EHRs. In doing so, our approach enables the review of predefined policies, as well as the discovery of unknown behaviors. We evaluate our approach with 5 months of access logs from the Vanderbilt University Medical Center and confirm the existence of stable social structures and intuitive business operations. Additionally, we demonstrate that there is significant turnover in the interactions between users in the HCO and that policies learned at the department-level afford greater stability over time.

1. Introduction

The healthcare community has made considerable strides in the development and adoption of information technologies [1]. Evidence indicates that the collection, storage, and processing of patient data in electronic form can decrease costs, strengthen staff productivity, and promote safety [2,3]. To realize these benefits on a broad scale, healthcare organizations (HCOs) must deploy information technology in a manner that facilitates business processes without jeopardizing patients' privacy rights [4]. HCOs are beginning to utilize software tools to design clinical information systems with logically sound privacy and security controls (e.g., [5–8]), however, such tools are limited in practice. In part, this is because they depend on HCO administrators to supply clear policy definitions, which has technical as well as social complications.

From a technical perspective, the specification of policies is a nontrivial challenge because a healthcare system is an inherently dynamic environment where teams of clinicians and support staff interact [9,10]. The notion of dynamic teams, while effective in supporting healthcare operations, is more complex than the pairwise relationship typically expected by access control frameworks (e.g., provider–patient) [11]. The complexity of the problem is compounded by the reality that HCOs evolve to address technological and organizational pressures, such as the adoption of new patient management protocols, reaction to legislated changes, and assimilation of rotating medical students [12].

From a social perspective, traditional methods for defining security policies are problematic because they rely on the knowledge of domain experts (e.g., employees) or the observations of external specialists. These methods may be feasible within the context of small systems, but are unmanageable for modern HCOs where the number of policies can be large, defined in an *ad hoc* manner, and revised at a moment's notice. The traditional approaches are further limited because they are subjective and can suffer from informant accuracy [13], which will be magnified by the scale of modern HCOs.

As an alternative, the growing adoption of electronic health record (EHR) systems provides an opportunity to apply automated learning methods that are data-driven, so that policies can be docked to actual behavior. Modern EHR systems already generate access logs to enable the construction of audit trails regarding what information care providers observe and what actions they take. Furthermore, many countries have enacted regulations that mandate their inclusion, such as the Security Rule of the US Health Insurance Portability and Accountability Act (HIPAA), which requires HCOs to retain access logs for a minimum of 6 years [14]. The automated extraction and representation of policies from such

* Corresponding author. Address: 2525 West End Avenue, Suite 600, Department of Biomedical Informatics, Vanderbilt University, Nashville, TN 37203, USA. Fax: +1 615 322 0502.
E-mail address: b.malin@vanderbilt.edu (B. Malin).

1532-0464/$ - see front matter © 2011 Elsevier Inc. All rights reserved.
doi:10.1016/j.jbi.2011.01.007

systems could complement traditional policy specification frameworks because they could assist in auditing existing policies, as well as discovering novel patterns of EHR system use. If such approaches can be designed in a scalable manner, they can endow administrators with a unique view of the stability, or volatility, of policies over time within their HCO.

We further anticipate that EHR user behavior can serve as the basis for automated surveillance tools that uncover policy violations. The problem could be framed as an anomaly discovery problem, such that accesses (or behaviors) that are sufficiently different than expected EHR system usage are flagged for administrative review. In this respect, we envision surveillance systems that mine policies from EHR access logs in a similar way to how intrusion detection systems mine rules from the audit logs of file systems [15,16]. However, a core assumption of such surveillance approaches is that behaviors manifest in regular patterns. As such, one of the primary goals of this paper is to assess if EHR system usage can be distilled into a representative set of patterns. While there are many different types of patterns that could be extracted, we limit our focus to patterns that summarize how EHR users and HCO departments collaborate. This is a fundamentally different problem than that studied in intrusion detection because EHR users are expected to function in a coordinated manner.

Specifically, we introduce a process to transform a database of EHR users' access of patient records into patterns of users' interactions, in the form of probabilistic rules. The process consists of two primary components; first we infer a social network of EHR users based on co-accesses to patient records. We focus on networks, so that we can capture the collaborative model of HCOs and their employees and, as we show, this network can be decomposed into clusters that represent expected organization structure at a high-level (e.g., children's vs. main hospital). Second, we convert the pairwise interactions of users into probabilistic rules of association. These rules capture the frequency of interaction between two users, as well as the likelihood of a user accessing a patient record given that another user accessed the same record. We focus on association rules because (as we review in Section 2) they have been shown to be effective at decomposing a complex environment into simple statistical components that can be applied in intrusion detection frameworks, but can also be presented to humans for review if need be. To make our approach reusable, we have developed open-source software to support the replication and application of our investigations.[1]

While there has been a number of investigations into the extraction of business process rules from access logs, as well as computational approaches for organizational modeling, to the best of our knowledge; this is the first research to link the concepts. Beyond a theoretical treatment of the problem, we perform an empirical analysis with 5 months of access logs from a large HCO's EHR system that is well-integrated into core healthcare operations. Through this investigation, we confirm that the HCO is a highly dynamic environment, which is not necessarily appropriate for traditional access policies. At the same time, we demonstrate that automatically learned policies can be grouped into levels of stability with intuitive implications.

2. Related work

Historically, automated learning approaches have proven to be capable of extracting relevant knowledge from access logs. Much of the work in log mining parallels the growth of the Internet, where it was shown that a domain's webpage access logs could be modeled as a database of transactions [17], from which association rules could be discovered to characterize users' tendencies [18,19]. Beyond log analysis methodology, it was demonstrated that knowledge, mined from access logs, can inform business decisions and can provide feedback to make systems more efficient. As a result, e-commerce sites routinely employ access log analytics to influence, attract, and retain customers [20–25]. Beyond marketing, there is some evidence in the healthcare domain which indicates that EHR access logs can reveal the clinical information use behavior for patients [26] and providers (e.g., in one study it was shown that care providers often viewed a patient's laboratory and radiology results in the same session) [27–29]. Further work has shown that logs are effective for understanding how medical information systems are used for educational purposes [30–32].

Nonetheless, research on access log analytics has, in general, focused on "what" users view, but HCO policies need to model "who" is viewing the health records of whom and "why". To address the latter, it helps to know the reason for a particular access; however, such information is not always adequately documented in an EHR. As such, in this research we focus on the extent that basic information in EHR access logs can provide intuition into the patterns of use that care providers generate when accessing patient records. In doing so, our goal is to uncover the relationships between individuals in the healthcare enterprise and the types of health record access workflows that exist. In this respect, methods from the social network analysis [33] and computational organizational science [34] domains are ideal candidates for policy extraction. In the biomedical realm, such techniques have been successfully applied to characterize the interactive and dynamic nature of multi-disciplinary research communities [35,36], to discover the dynamics of public health settings to facilitate change management [37], to explain physician adoption rates of EHR systems [38], and, most recently, investigate the relationships of healthcare providers in various clinical contexts [39–41].

In the context of health information security, we recognize that auditing through EHR access logs has become a practice both for HCOs and, more recently, for patients. One of the primary features that healthcare consumers desire in an EHR is transparency and, in particular, consumers want to be provided with the ability to access their medical records, as well as determine who has viewed them [42–44]. In support of this function, a system developed at New York Presbyterian Hospital was built [45] to enable the latter and found that such a tool can lessen the burden for EHR administrators by increasing the number of individuals that can monitor for violations. At the same time; however, it was observed that the task of analyzing the logs is complex due to the number of access that occurs for a typical patient's record, as well as the fact that patients are neither well-versed in an HCO's operations nor comprehend if a user has a legitimate reason to access their record.

To assist in EHR access logs monitoring, Asaro and Ries [46] prototyped a tool to automatically analyze EHR access logs by inspecting the distribution of records accessed by users. In their investigation, it was observed that some users accessed an excessive number of medical records; however, no methods were designed to determine if such behavior was in line with the behavior of the HCO. Thus, while the scientific literature exists for setup, collection, and summarization of EHR access logs, the study of the information within the access logs for health information security and policy evaluation has been neglected.

3. Relational healthcare policy discovery

This section describes the framework and methods applied for extracting healthcare policies from EHR access logs and begins with a high-level overview. The framework consists of two core

[1] A freely available open-source software tool can be downloaded from http://code.google.com/p/hornet/source.