Mach Learn (2014) 95:87–101
DOI 10.1007/s10994-013-5376-1

# Detecting inappropriate access to electronic health records using collaborative filtering

**Aditya Krishna Menon · Xiaoqian Jiang · Jihoon Kim · Jaideep Vaidya · Lucila Ohno-Machado**

Received: 16 November 2012 / Accepted: 29 April 2013 / Published online: 28 June 2013

**Abstract** Many healthcare facilities enforce security on their electronic health records (EHRs) through a corrective mechanism: some staff nominally have almost unrestricted access to the records, but there is a strict *ex post facto* audit process for inappropriate accesses, i.e., accesses that violate the facility's security and privacy policies. This process is inefficient, as each suspicious access has to be reviewed by a security expert, and is purely retrospective, as it occurs after damage may have been incurred. This motivates automated approaches based on machine learning using historical data. Previous attempts at such a system have successfully applied supervised learning models to this end, such as SVMs and logistic regression. While providing benefits over manual auditing, these approaches ignore the *identity* of the users and patients involved in a record access. Therefore, they cannot exploit the fact that a patient whose record was previously involved in a violation has an increased risk of being involved in a future violation. Motivated by this, in this paper, we propose a collaborative filtering inspired approach to predicting inappropriate accesses. Our solution integrates both *explicit* and *latent* features for staff and patients, the latter acting as

Editors: Kiri Wagstaff and Cynthia Rudin.

**Electronic supplementary material** The online version of this article (doi:10.1007/s10994-013-5376-1) contains supplementary material, which is available to authorized users.

A.K. Menon (✉) · X. Jiang · J. Kim · L. Ohno-Machado
UC San Diego, La Jolla, CA 92093, USA
e-mail: akmenon@ucsd.edu

X. Jiang
e-mail: x1jiang@ucsd.edu

J. Kim
e-mail: j5kim@ucsd.edu

L. Ohno-Machado
e-mail: machado@ucsd.edu

J. Vaidya
Rutgers University, Newark, NJ 07102-1897, USA
e-mail: jsvaidya@rbs.rutgers.edu

a personalized "fingerprint" based on historical access patterns. The proposed method, when applied to real EHR access data from two tertiary hospitals and a file-access dataset from Amazon, shows not only significantly improved performance compared to existing methods, but also provides insights as to what indicates an inappropriate access.

**Keywords** Access violation · Collaborative filtering · Electronic health records · Privacy breach detection

## 1 Problem of interest

In healthcare, the proliferation of computerized databases, inter-networking of systems, and search and indexing technology for electronic records have opened up several new possibilities and brought several advantages. To name a few, these include improved efficiency of record keeping, easier detection and prevention of fraud, waste, and abuse, and an increase in the overall quality of provided healthcare. However, the use of technology has also expanded the chances of inappropriate, unauthorized, or illegal access to (and use of) personal information (Office of Technology Assessment, United States Congress 1986). In the healthcare domain, recent breaches include the inappropriate viewing of the health records of celebrities (Ornstein 2008), a government official (Porter 2010), and a mother who gave birth to octuplets (Kaushik and Ramamurthy 2011). In 2011, 139 privacy breaches were reported, with each case affecting over 500 patients according to the notification mandated by the US Department of Health and Human Services (HHS) (Guardian 2010). The responsible institutions were fined and their reputations suffered serious damage. The completeness, relevance and sensitiveness of Electronic Health Records (EHR) make the healthcare privacy challenge even more important. A recent survey indicates that 71 % of physicians are concerned about health information exchange's effect on privacy (Wright et al. 2010) because security and privacy breaches can significantly undermine the patient's trust, imposing negatively on the quality of care.

Healthcare institutions, including the one participating in this study, have worked hard to keep their patients' records secure and private. Examples of adopted privacy protection methods include encrypted devices, strong passwords, two-factor authentication, employee training on privacy, annual signing of confidentiality agreements, and self-audit tools for employees to see who has viewed their records (Boxwala et al. 2011). Despite these safeguards, a continuing challenge is to be able to correctly classify whether a requested EHR access is in inappropriate or not, so that the former type can be blocked in advance. One approach to this is granting limited accesses to patient records for the purpose of treatment, payment, healthcare operation and research. However, it is difficult to impose a static predefined access control policy on employees in this setting since the hospital work environment has dynamic and unpredictable care patterns, and flexible workflows where providers are assembled unexpectedly (such as in the emergency room). The continuing trend towards a mobile workforce and team-based care (including teaching, and ever-changing job titles and roles) further adds to this complexity (Boxwala et al. 2011).

These challenges motivate the use of automated systems to at least partially alleviate the burden of determining inappropriate accesses. One can aim to train a machine learning model on historical access data, and use it to classify the appropriateness of future accesses. Prior work in this direction has used supervised learning techniques on certain features of the staff members and patients involved (e.g. Boxwala et al. 2011; Kim et al. 2011; see Sect. 3.3 for more discussion). These papers have demonstrated the potential of using machine learning models to address privacy issues in EHRs.

In this paper, we look to extend these prior machine learning approaches to exploit an additional, important source of information: the *identity* of the user and patient involved in the access. This lets us capture the fact that, for example, a user who has performed a violation in the past is more likely to commit a violation in the future. To this end, we propose a *collaborative filtering* inspired approach to create a more fine-grained model for detecting inappropriate accesses. Our model integrates the traditional features for users and patients with *latent features* that can be thought of as a "fingerprint" based on their historical access behavior. We show how such a model can be adapted to meet the characteristics of this problem; specifically, we show how to *pool together* data for reliable parameter estimation, and how to create *interaction-specific* predictions. Our proposed method, when applied to real EHR access data from tertiary hospitals and a file-access dataset from Amazon, shows significantly improved performance compared to existing methods, and provides insights as to what indicates an inappropriate access.

Formally, this paper focuses on detecting privacy breaches resulting from inappropriate accesses of EHRs. We call this the *access violation* problem. As input, we are given a labelled training set $\mathcal{T} = \{(a^{(i)}, y^{(i)})\}_{i=1}^{N}$, where $a^{(i)}$ comprises features that summarize an access to a patient's EHR by a hospital employee or its affiliated member (who we shall call a *user*) that is logged in to a healthcare system, $y^{(i)} \in \{0, 1\}$ is a label denoting whether the access was a violation of hospital policy or not, and $N$ is the number of labelled instances. Generally, the features in $a^{(i)}$ can be broken down as follows:

(i) *identity* features, namely, the user's identifier $u \in \{1, \dots, m\}$, as well as the patient's identifier $p \in \{1, \dots, n\}$, where $m$ and $n$ indicate the total number of users and patients in the system, respectively,
(ii) *user-specific* features $x_u \in \mathbb{R}^{d_1}$ corresponding to the user performing the access (e.g. is a physician? has an administrative job role?),
(iii) *patient-specific* features $x_p \in \mathbb{R}^{d_2}$ corresponding to the patient whose record is being accessed (e.g. had a clinic visit within a week? is a VIP patient?), and
(iv) *relational* features $x_r \in \mathbb{R}^{d_3}$ specific to the access (e.g. do the user and patient share the same last name?).

Our task is: given a new access $a' = (u', p', x'_u, x'_p, x'_r)$, can we predict whether or not it is a violation?

## 2 Data preparation

In Sect. 1, we described the data arising in access violation problems in abstract terms. In practice, there are several important characteristics of the data that impact performance and model design. First, there are relatively few labelled training examples. Collecting labelled data is a daunting task for security officers, because they provide a lengthy, in-depth analysis of each case and are therefore generally unable to look at more than a few hundred cases per review period. Second, access violations are generally rare, as most employees self-moderate. Thus, amongst the labelled examples, one would expect the classes to be highly imbalanced, with a small fraction of examples being labelled as violations. Third, the data likely suffers from sample-selection bias. Indeed, we found that security officers tend to provide labels for conspicuously improper accesses, which may be unrepresentative of future accesses. (We discuss another manifestation of this bias in Sect. 4.2.)

Our experiments are based on two real-world datasets. The first is data collected for a prior study of machine learning techniques for privacy breach detection (Kim et al. 2011). We call this dataset "Hospital". The raw data comprised 34.1 million EHR accesses from the operational databases of two tertiary hospitals in Boston, MA, USA over six months in the year 2009. From this large pool, security experts provided 1504 accesses a positive or negative label, denoting a violation or safe access respectively. A further 4958 accesses were annotated with features, but not labelled; these comprise unlabelled examples that one may hope to use for learning. In total, there are 313 users and 319 patients in the labelled set, with a positive to negative label ratio of about 1 : 7. We have 9 event-side features, 7 user-side features, 12 patient-side features and 9 relational features. The full list of feature names and their detailed descriptions can be found in Boxwala et al. (2011). Given the sensitive nature of the data, we are unfortunately unable to release it for public use.

The second dataset is the recently released Amazon Access data,[1] which involves the task of predicting whether users are allowed to access a file or not. We call this dataset “Amazon”. While the semantic meaning of the data is not exactly the same to the EHR access scenario that is the focus of this paper, there are enough similarities in the goal to justify using the same model for this scenario. We include results on this dataset since it is publicly available, and thus serves as a potential benchmark for future research. The dataset comprises 1 027 347 labelled accesses of 79 906 users on 11 615 files. The accesses are either labelled as positive, denoting access is allowed, or negative, denoting access is denied or revoked. Each user also has 11 categorical features denoting the manager that he reports to, his current role in the company, and so on.

## 3 Machine learning technique

We now discuss the motivation and mechanics of our machine learning model.

### 3.1 A collaborative filtering model

We analyze the access violation problem in terms of *dyadic prediction* (Hofmann et al. 1999), where the goal is to predict a label for the interaction of a pair of entities. This framework captures scenarios such as recommending friends in a social network (Yang et al. 2011), predicting student performance on test scores (Thai-Nghe et al. 2011), and click-through rate prediction in computational advertising (Menon et al. 2011). A well-studied instance of dyadic prediction is the task of predicting users’ ratings for items based on their past preferences. *Collaborative filtering* is a popular strategy for this problem, where one attempts to tease out the implicit characteristics of users and items based solely on these historical preferences, and use these to predict preferences for every (user, item) pair. Arguably the most powerful collaborative filtering approach is the *latent feature* model (Koren et al. 2009), where the predicted label for the pair $(i, j)$ is

$$\hat{y}\big((i, j); \theta\big) = f\left(\alpha_i^T \beta_j + \gamma_i + \delta_j + \mu\right)$$

for some appropriate link function $f(\cdot)$. The terms here include a global bias $\mu$, user- and item-specific biases $\gamma_i \in \mathbb{R}, \delta_j \in \mathbb{R}$, as well as an interaction term $\alpha_i{}^T\beta_j$, where $\alpha \in \mathbb{R}^{k\times m}$, $\beta \in \mathbb{R}^{k\times n}$ for some number $k \in \mathbb{Z}^+$ of *latent features*.

Observe that our training set $\mathcal{T}$ may be encoded as an $m \times n$ table[2] $Y$, where $m$ is the number of users, $n$ the number of patients, and each cell $(i, j)$ collects all outcomes of the

[1] https://sites.google.com/site/amazonaccessdatacompetition/.

[2] If each user accessed each patient record at most once, this table could be represented as a matrix.

user $i$ accessing the patient $j$'s record:

$$Y_{ij} = \{(a, y) : (a, y) \in \mathcal{T} \wedge u(a) = i \wedge p(a) = j\}.$$

This is similar to an item recommendation problem: we are trying to measure the affinity a particular (user, patient) pair has for violation. The "rating" that a user gives a patient is a measure of the appropriateness of the corresponding interaction. (Unlike classical item recommendation, here, each cell can comprise *multiple* entries, because a user can access the same patient's record multiple times, in different contexts.) With this interpretation, we may look to apply techniques from collaborative filtering to the access violation problem. In particular, a latent feature approach models the suspiciousness of the access $a$ as

$$\hat{y}(a; \theta) = f\left(w^T \phi(a) + \alpha_u^T \beta_p + \gamma_u + \delta_p + \mu\right).$$

We can interpret $\alpha_u$ and $\beta_p$ as comprising certain *latent features* of the user and patient respectively. We additionally allow for a term $w^T \phi(a)$ that leverages information present in any *explicit features* $\phi(a)$ for the access.

This model has several desirable characteristics. First, it captures strictly more information than a supervised learning method, by virtue of augmenting explicit features with latent features that are learned from data. Second, it may be trained efficiently on large datasets using stochastic gradient descent (Koren et al. 2009). Third, stochastic gradient training allows for incremental updates to the learned model, rather than full retraining. Fourth, such models have been shown to work well in traditional collaborative filtering applications, so one may hope that they perform similarly well in this domain.

### 3.2 Extending the basic collaborative filtering model

We now discuss some domain-specific extensions to the collaborative filtering model. (We defer the mathematical details to the Supplementary Material.)

One basic issue with access violation data is the scarcity of labels, since each label must be determined by security experts. Consequently, most rows and columns in the (user, patient) table will have only a few labels. How do we reliably estimate latent features from such data? To meet this challenge, we use the intuition that while at the user- and patient-level accesses are sparse, they are denser if we *pool together* users and patients into *clusters* or *bins*. We may then estimate *coarse* latent features at this cluster level, and use this as a prior for the more personalized user- and patient- latent features. Intuitively, all characteristics being equal, we would expect the access patterns of two cardiologists to be similar, and so given an estimate for a general cardiologist's latent vector, we can push every individual cardiologist's vector towards this. An appealing consequence of this pooling process is that it gives a principled way of handling the *cold start* problem, namely, new users or patients for whom we do not already have any labelled accesses. One certainly expects there to be a constant flow of new patients to any healthcare facility, making this a pertinent problem for this domain. Our approach effectively lets such cold-start entities inherit the characteristics of all other (labelled) entities in the same cluster.

Another issue is that the latent feature model above keeps a context-independent score of user $u$ accessing patient $p$, via the term $\alpha_u^T \beta_p$. This ignores the contextual information present in the relational feature $x_r$; this is only taken into account by the weights on the interaction features in the $w^T \phi(a)$ term. While sensible, one may hope to directly estimate *context-dependent* latent features. For example, we could use a separate set of latent vectors depending on whether the relation involves family members, or not: a user may be scrupulous in general, but unreliable when it comes to people he is related to. We can thus modify the model to estimate a separate set of latent features for each possible interaction context.

### 3.3 Comparison to existing work

Recently, much work has been done on automated detection of inappropriate access to health records, and the related problem of determining if a user conducts anomalous accesses. In Chen and Malin (2011), the problem of detecting anomalous users was approached by inferring a set of social structures, on top of which anomaly detection was applied. In Zhang et al. (2011), classifier-based role prediction was performed to assign a limited access to a user. Association rule mining on an inferred social network of EHR users based on access patterns was used in Malin et al. (2011). Another framework was developed in Fabbri and LeFevre (2011) to automatically extract a reason for each EHR access, which may be practically useful. Unfortunately, these studies did not use labelled EHR access in a real hospital setting, and therefore did not report performance metrics of real access violation detection (e.g., 0-1 accuracy or area under the ROC curve). In contrast, we use labelled samples from real EHR accesses, based on data collected for the study (Boxwala et al. 2011). In the latter study, acting security officers in the participating institutions provided labels, and this was expanded by requesting more labels on selected samples using active learning. The initial study performed classification using logistic regression, and this was later extended to use clustering and rule-based detection (Boxwala et al. 2011). Considerable efforts were spent on designing predictive features.

The key difference to Boxwala et al. (2011) and Kim et al. (2011) is that we move from supervised learning to collaborative filtering. A pure supervised learning approach has two limitations. First, its efficacy relies on the underlying features being predictive of whether or not the access is likely to be a violation. But specification of such explicit features is difficult, and might not be comprehensive. Second, it ignores the identity of the user and patient for a single access, which may be sub-optimal. Suppose that there is a particular patient whose record has been illegally accessed in the past. Then, it is reasonable to elevate the suspicion of any future access of this patient's record being a violation, because for example it may be the case that this patient has some "interesting" condition (e.g., the aforementioned mother who gave birth to octuplets). One way to overcome this limitation is to augment the feature representation $\phi(a)$ to include information about the access history of the user $u$ and patient $p$. However, this strategy will always end up playing "catch up" to find a suitable representation of prior accesses. The collaborative filtering approach aims to automatically uncover these complex relationships.

## 4 Experimental design

We now discuss the design of our experiments, and some dataset processing details.

### 4.1 Methods compared and evaluation scheme

We compared the basic collaborative filtering model using logistic loss ("CF")[3] with three supervised learning methods, namely linear regression, logistic regression, and SVM with a linear kernel.[4] For the supervised learning methods, we experimented with both using the raw feature set, as well as augmenting it with features derived from past history

[3] We observed qualitatively similar results for the square loss.

[4] We did not experiment with nonlinear versions of these methods. Adding nonlinearity would likely improve performance, but would still be limited by the fact that no identity information is used when learning. Thus, as argued in Sect. 3.3, there is a fundamental limitation to supervised learning as compared to collaborative filtering.

("+History"), such as the number of prior violations the user/patient was involved in and the fraction of accesses that were violations,. For the collaborative filtering model, we used the extensions described in this paper, where appropriate progressively adding explicit features ("+Feats"), pooling together latent features ("+Pooling"), and having relational feature components ("+Rel"). Our final model used all components in conjunction.

We evaluated performance of various methods using nested 5-fold cross-validation, where each training fold was split into 2 further folds for hyperparameter selection. Our performance metrics are the root mean squared error (RMSE) of the model predictions to the access label, the area under the ROC (AUROC) (Fawcett 2006) and precision-recall (AUPRC) curves (Davis and Goadrich 2006) (where the curves plot the false positive vs true positive rate and precision vs recall respectively), and the geometric mean of the sensitivity and specificity of the classifier (Kubat and Matwin 1997) with a default threshold of 0.5 (G-Mean). The latter three metrics are generally employed when classes are imbalanced, and/or when there is an implicit, but unknown, cost-sensitive nature to the incorrect predictions. The RMSE measures the fidelity of model predictions as probability estimates, which is useful if we were to take downstream actions based on the confidence score outputted by the classifier (for example, selecting the accesses most confident of being (non-)violations, and sending them to a security officer for a time-consuming review.).

### 4.2 Processing of the datasets

The Hospital data has a quirk that forced us to process it before applying collaborative filtering methods: namely, for each user, either all of her accesses are violations, or none of them are. The reason for this is the data was collected for the study of Boxwala et al. (2011), where user and patient identifiers were not factored into the learning algorithm; as security officers were not directed otherwise, they found it natural to identify suspicious users, and collect a random sample of their accesses, rather than directly collect a random sample of suspicious accesses. As a result of this issue, a collaborative filtering model will simply memorize this "trust-worthiness" of each user as part of the bias term, and appear to be perfectly predictive of violations. This issue forced us to *remove* the user and patient identifiers when training our collaborative filtering models. However, we retained the *cluster* or *bin* for each user and patient, as described in the previous section. Thus, while we no longer know the raw identities of users and patients, we may still operate on the user *cluster* by patient *cluster* matrix.

We clarify some points regarding the above processing. First, it has *no* influence on the supervised learning baselines as studied in Boxwala et al. (2011) and Kim et al. (2011), as it only modifies information about the user and patient identifiers, which these methods do not exploit. Second, the processing only *hampers* collaborative filtering methods, by allowing us to use strictly less information than we would like in a real world usage of such a system. As we shall see, even this handicapped version of collaborative filtering performs favorably compared to existing baselines. Third, it only applies to the Hospital dataset; as there is no similar issue in the Amazon dataset, we use it as-is, and apply all previously described variants of our method.

## 5 Empirical results

We now present results that aim to determine whether our collaborative filtering model allows us to catch more privacy breaches than existing methods.

**Table 1** Cross-validation performance with standard deviations, Hospital dataset

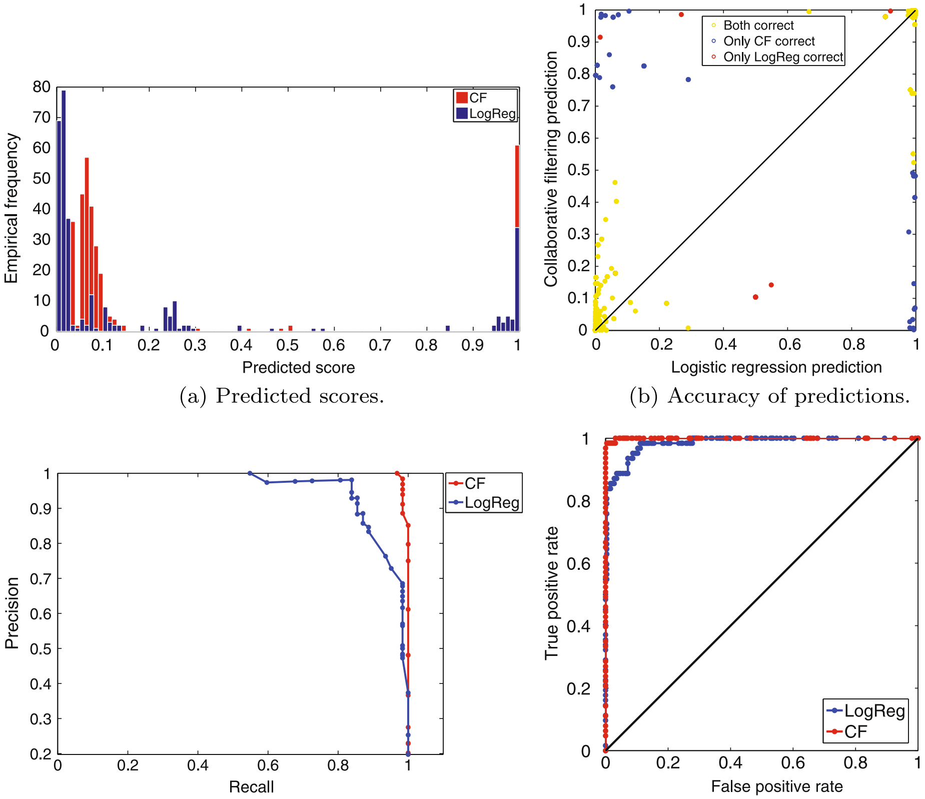
| Model | RMSE | AUROC | AUPRC | G-Mean |
|---|---|---|---|---|
| Linear reg. | $0.2139 \pm 0.0217$ | $0.9642 \pm 0.0227$ | $0.8838 \pm 0.0492$ | $0.8767 \pm 0.0388$ |
| Logistic reg. | $0.2049 \pm 0.0248$ | $0.9688 \pm 0.0175$ | $0.8935 \pm 0.0459$ | $0.8876 \pm 0.0401$ |
| SVM | $0.2875 \pm 0.0155$ | $0.9658 \pm 0.0231$ | $0.8938 \pm 0.0463$ | $0.8834 \pm 0.0382$ |
| Linear reg. + History | $0.2121 \pm 0.0216$ | $0.9651 \pm 0.0213$ | $0.8871 \pm 0.0491$ | $0.8757 \pm 0.0397$ |
| Logistic reg. + History | $0.1957 \pm 0.0252$ | $0.9800 \pm 0.0105$ | $0.9101 \pm 0.0443$ | $0.8919 \pm 0.0374$ |
| SVM + History | $0.2382 \pm 0.0311$ | $0.9779 \pm 0.0109$ | $0.9045 \pm 0.0432$ | $0.8885 \pm 0.0402$ |
| CF | $0.1634 \pm 0.0269$ | $0.9900 \pm 0.0055$ | $0.9368 \pm 0.0457$ | $0.9251 \pm 0.0314$ |
| CF + Feats | $0.1398 \pm 0.0307$ | $0.9924 \pm 0.0067$ | $0.9687 \pm 0.0248$ | $0.9456 \pm 0.0309$ |
| CF + Rel | $0.1510 \pm 0.0305$ | $0.9921 \pm 0.0051$ | $0.9578 \pm 0.0264$ | $0.9427 \pm 0.0319$ |
| CF + Feats + Rel | $0.1380 \pm 0.0310$ | $0.9916 \pm 0.0118$ | $0.9714 \pm 0.0275$ | $0.9402 \pm 0.0329$ |

### 5.1 Results on Hospital data

Table 1 summarizes the results on the Hospital dataset. We draw the following conclusions from our results:

- Linear and logistic regression are reasonably strong baselines, and perform favorably on all performance metrics.
- On all metrics, the basic collaborative filtering model (that only uses user and patient cluster identifiers) performs at least as well as the explicit feature baselines of linear and logistic regression. This indicates that there is strong signal present in the user and patient cluster identifiers, and that they are largely as informative as even interaction-specific features.
- Taking into account historical accesses in the supervised learning models improves their performance. However, the CF model significantly outperforms this extended feature representation. This says both that taking into account identities and past history can give an important benefit over the information contained in the features alone, and that the influence this information has on the label is potentially complex.
- With respect to all metrics, combining latent and explicit features at least maintains, and at best improves the overall performance of either individual model. This indicates that there is potentially complementary information in the (cluster-level) user and patient identifiers over just the explicit features.

To get a clearer sense of the distinction to logistic regression, we plot in Fig. 1(a) the histograms of scores produced by both models. The clear differences are that the collaborative filtering model predicts more examples to be highly likely of violation, and that it is more conservative about predicting an example to not be a violation. In Fig. 1(b), we plot the predicted scores of logistic regression and our collaborative filtering model on all 1 504 examples. For each example, we determine which of the two methods makes a correct prediction (after thresholding the scores at 0.5), and colour the points in the scatterplot appropriately. We see that there are several cases where logistic regression is incorrectly overconfident (i.e., makes a prediction close to 0 or 1), while the collaborative filtering model is more conservative on the correct side on the threshold. (Note that thresholding to give a binary score is a simplification of the actual process followed based on a classifiers' predictions, which will likely exploit the confidence values in some domain-specific manner. It is here that better RMSE's, indicating better probability modeling, become important.)



(a) Predicted scores.

(b) Accuracy of predictions.

(c) Precision-recall curves.

(d) ROC curves.

**Fig. 1** Comparison of models on Hospital dataset

The results for the Hospital dataset suggest improvements of between 2 %–4 % in the ranking metrics. Given the relatively high scores of the baselines, however, it is prudent to attempt to quantify the practical significance of the improvements. Figures 1(c) and 1(d) study the precision-recall and ROC curves of logistic regression versus our collaborative filtering model. We see that the latter achieves perfect precision with a recall nearly twice that of logistic regression. This increased detection rate is important, given the cost associated with these violations. The ROC curve shows that the collaborative filtering model almost perfectly separates violations from non-violations, and in particular manages to reduce the number of false positives over logistic regression for a range of thresholds. At almost perfect positive recall, the collaborative filtering model cuts down the false positive rate by over a factor of four. We further observed that the collaborative filtering curves strictly dominate those of logistic regression, indicating that for every choice of threshold, we have a better false and true positive rate.

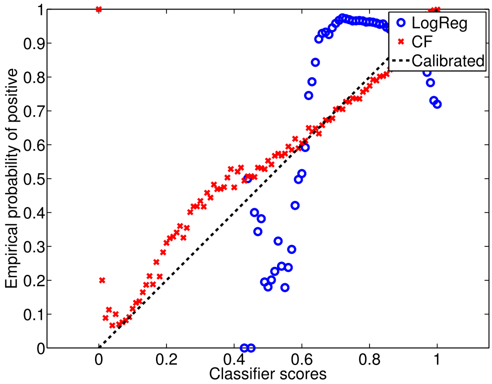
## 5.2 Results on Amazon data

Table 2 summarizes the results on the Amazon dataset. (As the dataset does not include (user, file) features, it is not possible to evaluate the efficacy of learning relation-specific features here.) We make the following observations:

**Table 2** Cross-validation performance with standard deviations, Amazon dataset

| Model | RMSE | AUROC | AUPRC | G-Mean |
|---|---|---|---|---|
| Linear reg. | 0.2724 ± 0.0307 | 0.6198 ± 0.0231 | 0.9559 ± 0.0027 | 0.0920 ± 0.0339 |
| Logistic reg. | 0.2290 ± 0.0010 | 0.7655 ± 0.0021 | 0.9784 ± 0.0004 | 0.0677 ± 0.0500 |
| SVM | 0.3295 ± 0.0002 | 0.7712 ± 0.0029 | 0.9787 ± 0.0006 | 0.0277 ± 0.0143 |
| CF | 0.2213 ± 0.0008 | 0.9102 ± 0.0009 | 0.9931 ± 0.0001 | 0.4622 ± 0.0046 |
| CF + Feats | 0.2172 ± 0.0009 | 0.9133 ± 0.0013 | 0.9934 ± 0.0001 | 0.4715 ± 0.0095 |
| CF + Pooling | 0.2177 ± 0.0010 | 0.9102 ± 0.0007 | 0.9931 ± 0.0001 | 0.4629 ± 0.0051 |
| CF + Pooling + Feats | 0.2084 ± 0.0009 | 0.9189 ± 0.0009 | 0.9938 ± 0.0001 | 0.5024 ± 0.0090 |

**Fig. 2** Reliability diagram on Amazon dataset



- Explicit features do not seem very helpful on this dataset. This is likely a consequence of the fact that they are very sparse in instance space, that is, each feature is only active for a small fraction of examples. (Indeed, the median fraction of examples where a feature is on is ∼ 0.02 %.)
- Collaborative filtering offers a significant improvement over supervised learning methods in this dataset. The area under the ROC curve, for example, increases by over 15 %. This is in contrast to the Hospital dataset, where the improvements were much more modest.
- Our latent-feature pooling method improves performance slightly, but reliably. We note that the final model that employs pooling and explicit features sees significant improves in both the RMSE and G-Mean metrics, though the rankings stay essentially the same. The improvement in RMSE suggests that the quality of the probability estimates found by the model are superior to the supervised learning methods.

The large differences between our method and the supervised learning baselines in the ranking metrics lead one to expect similar dominance in the underlying curves. Indeed, this is the case, but due to space constraints, we defer these plots to the Supplementary Material. To analyze the results on a different dimension, Fig. 2 plots a reliability diagram (Murphy and Winkler 1977) for the two models. The purpose of this diagram is to evaluate the fidelity of the probability estimates returned by a model. The $x$-axis comprises classifier scores, and for a given score, the $y$-axis comprises the fraction of positives amongst all examples that are predicted to have that score. (For clarity, such a diagram is generally produced by binning the classifier scores.) A model for which the curve is a diagonal is said to be *calibrated* (DeGroot and Fienberg 1983). The curve shows that the collaborative filtering model generally produces a tighter fit to the diagonal, indicating it provides better probability estimates. Logistic regression appears to especially suffer for predictions that are close to 1: we see that there is a sudden dip below the diagonal after the 0.5 threshold, which corroborates our explanation for the non-monotonicity of the precision-recall curve.

## 6 Expert commentary and potential infusion

We comment on the practical viability of the model proposed in this paper, and discuss challenges that it (and any other machine learning model) must overcome to truly provide an automated solution to the access violation problem.

### 6.1 Potential real world impact

Towards the end of the study in Boxwala et al. (2011), the participating hospital system conducted investigations based on the learning model's predictions, and imposed sanctions on offending users when needed (Kim et al. 2011). In fact, in one case a user faced termination of employment. Other offenders were given warnings and were re-educated about the institution policies. The institution also conducted an employee education campaign, reinforcing the "Minimum Necessary" access policy. The campaign enlightened staff who were not aware of certain aspects of hospital policy, for example that family member accesses are forbidden.

The above makes clear that machine learning systems can have a high impact in this domain. Given that our model was designed to extend the successful supervised learning models, and that it showed good empirical performance, it suggests that it may facilitate tangible real-world value. As a further benefit, the proposed machine learning system is devised to detect rare events in large-scale hospital data. So, the same method could potentially be applied to other rare event detection problems in healthcare, such as detecting post-operative complications, adverse drug events, medical insurance fraud, *et cetera* (Boxwala et al. 2011; Kim et al. 2011). However, as we now discuss, simply performing better than existing models is not sufficient to constitute a practically significant advance in the domain.

### 6.2 Why accuracy is not enough: gaining insights from the model

From a machine learning perspective, our proposed algorithm is novel for this domain, and our experimental results showed that it can be significantly more accurate in discriminating between appropriate and inappropriate accesses. However, from the domain experts' perspective, these improvements were not seen as especially appealing. The machine learning model is largely treated as a black box, and so modifying this box to produce better results was of course appreciated, but was not as exciting as going from manual to automated auditing. However, what *did* appeal was the ability to gain new qualitative insights about the data. As our collaborative filtering model infers latent features for users and patients, unlike supervised learning models, we can hope to gain different insights compared to the supervised learning models.

To attempt to gain insights from the model, we added a visualization layer summarizing model-generated prediction results to enable the security officer's decision support in an intuitive manner. We constructed a user-patient interaction map that reveals which subgroups of users and patients have a high risk of being involved in an inappropriate access. Figure 3 displays the user-based patterns in columns and patient-based patterns in rows. The color

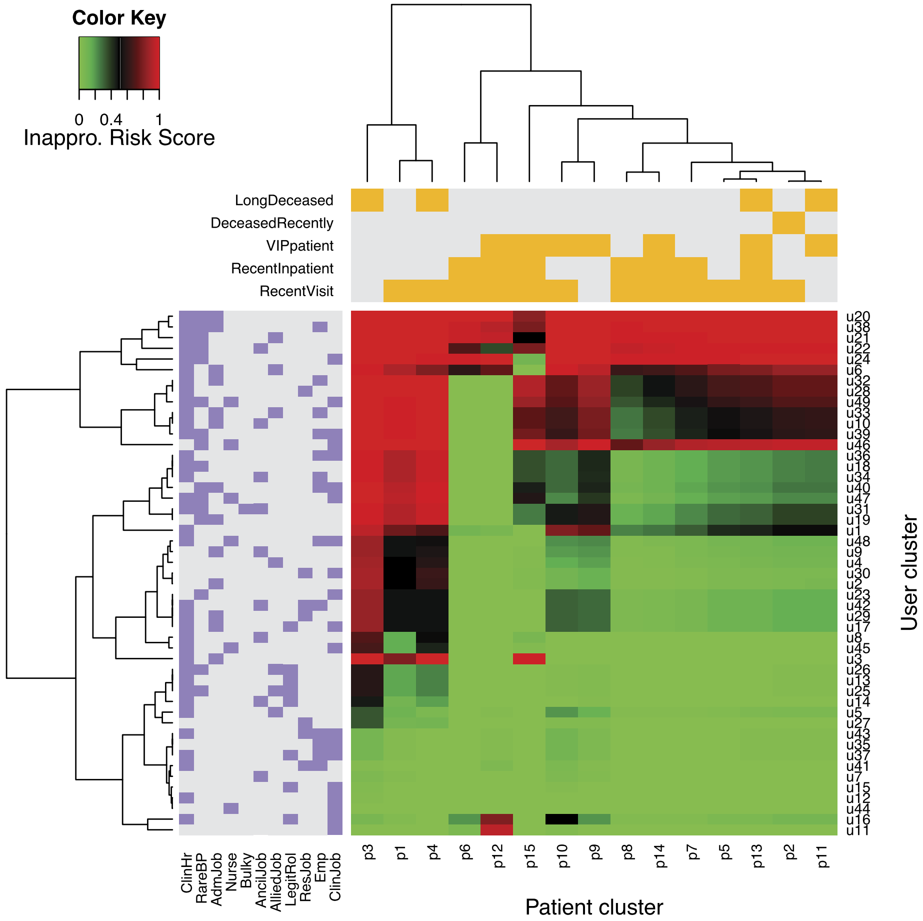


**Fig. 3** Heatmap for user-patient interaction. Each cell represents a prediction score for inappropriate access ranging from 0 (*green*) to 1 (*red*) (Color figure online)

of each cell represents the model's predicted score of an access involving the appropriate groups as being a violation (1) or a non-violation (0). The heat map, as a byproduct of collaborative filtering model, provides explainability and actionable findings, which can serve as a visual tool for evaluating employee integrity and measuring inappropriateness risk of the patients. For example, an access by user cluster 'U31', a bulky user who accesses more than 200 EHRs per day, is likely to be inappropriate if the patient being viewed has been deceased more than a year, so termed long-deceased. While the accesses from the user cluster 'U11' in the last row are mostly appropriate as these users have clinical jobs, an exception is made when a patient is VIP and has been hospitalized recently. Another simple lesson from the heatmap is that patients who have been deceased more than a year (first column) are highly likely to be victims of inappropriate EHR access.

As further analysis of the model, Fig. 4 visualizes the similarities between various user and patient groups in the learned latent space. For this visualization, we only keep the edges between groups that have weights in the 90th percentile; all other edges are removed, as are the resulting isolated nodes. The figure reveals some interesting structure about the data,

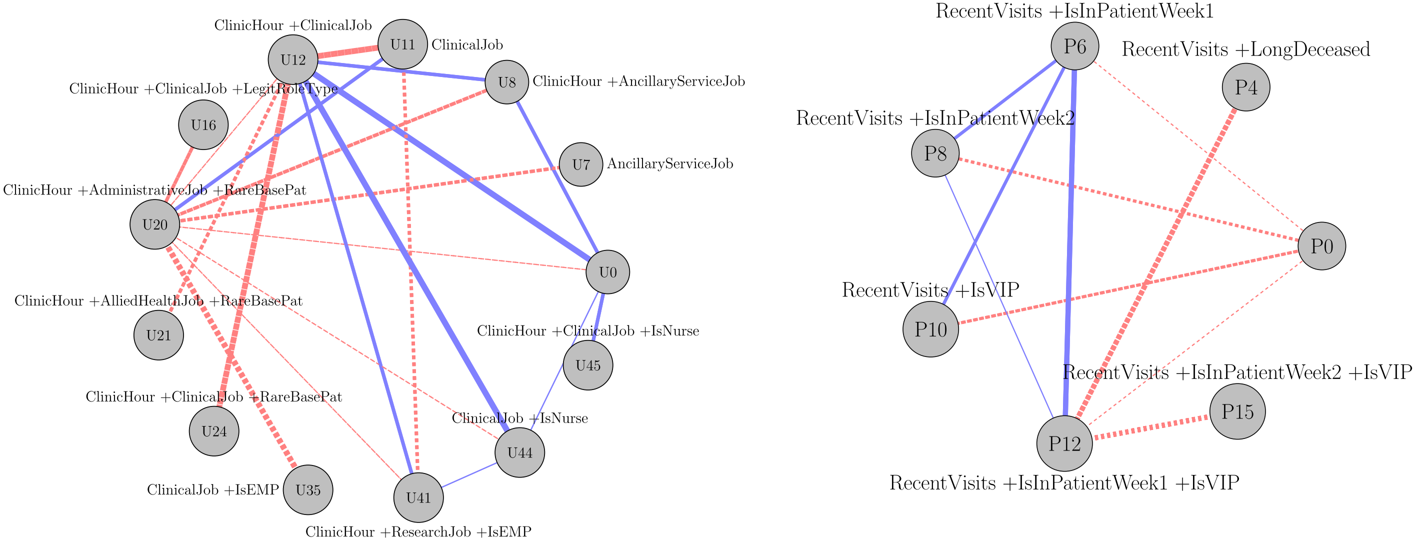

ClinicHour +ClinicalJob
U12
U11
ClinicalJob
U8
ClinicHour +AncillaryServiceJob
ClinicHour +ClinicalJob +LegitRoleType
U16
U7
AncillaryServiceJob
ClinicHour +AdministrativeJob +RareBasePat
U20
U0
ClinicHour +AlliedHealthJob +RareBasePat
U21
ClinicHour +ClinicalJob +IsNurse
U45
ClinicHour +ClinicalJob +RareBasePat
U24
ClinicalJob +IsNurse
U44
ClinicalJob +IsEMP
U35
U41
ClinicHour +ResearchJob +IsEMP
RecentVisits +IsInPatientWeek1
P6
RecentVisits +LongDeceased
P4
RecentVisits +IsInPatientWeek2
P8
P0
RecentVisits +IsVIP
P10
RecentVisits +IsInPatientWeek2 +IsVIP
P15
P12
RecentVisits +IsInPatientWeek1 +IsVIP

**Fig. 4** Similarities between user (*left*) and patient (*right*) groups in latent space. *Blue* (*solid*) edges denote similar groups, *red* (*dashed*) edges denote dissimilar groups. Thickness denotes strength of the tie

for example that VIPs who have recently visited the hospital have different access patterns depending on whether they were admitted one or two weeks ago (P12 to P15). Another observation is that users with clinical jobs who perform an access during a clinical hour have a different pattern of access to those who do it outside of a clinical hour (U11 to U12). These visualization tools promoted interaction and communication between ML researchers and security officers.

### 6.3 The human "bottleneck"

At the outset, we pointed out that the ideal end-system would be fully automated, and obviate the need for manual human effort. While previous works and our model take a step in this direction, they are far from a fully automated access violation system. The reason is that they involve collecting labelled examples from a security officer. For the study in Boxwala et al. (2011), the officers who provided labels had to go through the time-consuming process of multiple phone calls, emails, interviewing the suspect's supervisor, examining records in multiple disparate systems, *et cetera*. For this reason, the fraction of labeled instances in the Hospital dataset is relatively low (less than 0.01 %).

It is difficult in general to overcome the labelling "bottleneck", but there are at least two ways that machine learning models can attempt to mitigate this. First, the model can be used in an active learning framework, so that its predictions are used to inform which set of examples are sent to the security officers for review and labelling. Second, the models can try to exploit the vast amount of unlabelled data that is present. In preliminary experiments with the latter, however, we found that it was difficult to gain a significant improvement in performance.

## 7 Lessons for ML community

We believe our study has the following take-home messages for the ML community at large. First, pressing problems in healthcare may be modelled using techniques developed in the ML community for apparently disparate problems. Current systems for these problems tend to involve human classification, or relatively elementary ML models. Thus, there is scope for high impact by carefully adapting mature ML techniques to this domain. Second, close collaboration with end-users is required to ensure that the data collected matches what makes most sense for the model. In our Hospital data, we faced the issue of all labels associated with a particular user either being appropriate or inappropriate. This arose because the data was collected at a time when user and patient identifiers were not used as part of the modelling strategy: therefore, significant care is needed at the modelling stage to ensure that the effort expended in collecting data is not wasted if the model is changed slightly. Third, improving predictive performance, even significantly, is not always the biggest priority for the end-user. Our collaborators in the hospital system saw the value of ML as the decision support guiding tool, directing the security officers toward the cases with the most potential for being inappropriate, then using human resources in a smarter way to spend their limited time to investigate where the risk is highest. Fourth, there is need to design techniques that can effectively learn from a minimal number of labelled examples, as these are difficult to collect.

Overall, we believe there is scope to use ideas from social network analysis and collaborative filtering to improve detection of inappropriate accesses, and that our model is one step in this direction.

**Acknowledgements** The authors were funded in part by the NIH grants K99LM011392, R01LM009520, U54HL108460, R01HS019913, and UL1RR031980. We thank Janice M. Grillo, Rose Mandelbaum, Debra Mikels, Bhakti Patel and Partners HealthCare for their contributions to this research. We also thank the anonymous reviewers for their valuable comments that improved the presentation of the paper.

## References

Boxwala, A., Kim, J., Grillo, J., & Ohno-Machado, L. (2011). Using statistical and machine learning to help institutions detect suspicious access to electronic health records. *Journal of the American Medical Informatics Association*, *18*(4), 498–505.

Chen, Y., & Malin, B. (2011). Detection of anomalous insiders in collaborative environments via relational analysis of access logs. In *Proceedings of the first ACM conference on data and application security and privacy* (pp. 63–74). New York: ACM.

Davis, J., & Goadrich, M. (2006). The relationship between precision-recall and roc curves. In *Proceedings of the 23rd international conference on Machine learning, ICML'06* (pp. 233–240). New York: ACM. doi:10.1145/1143844.1143874

DeGroot, M. H., & Fienberg, S. E. (1983). The comparison and evaluation of forecasters. *Journal of the Royal Statistical Society. Series D. The Statistician*, *32*(1/2), 12–22. http://www.jstor.org/stable/2987588.

Fabbri, D., & LeFevre, K. (2011). Explanation-based auditing. *Proceedings of the VLDB Endowment*, *5*(1), 1–12.

Fawcett, T. (2006). An introduction to roc analysis. *Pattern Recognition Letters*, *27*(8), 861–874. doi:10.1016/j.patrec.2005.10.010.

Guardian, T. (2010). *Department of health & human services, breaches affecting 500 or more individuals*. Available online: http://www.guardian.co.uk/commentisfree/henryporter/2010/mar/02/nhs-spine-database-opting-out.

Hofmann, T., Puzicha, J., & Jordan, M. I. (1999). Learning from dyadic data. In *NIPS'99* (pp. 466–472).

Kaushik, R., & Ramamurthy, R. (2011). Whodunit: an auditing tool for detecting data breaches. *Proceedings of the VLDB Endowment*, *4*(12), 1410–1413.

Kim, J., Grillo, J., Boxwala, A., Jiang, X., Mandelbaum, R., Patel, B., Mikels, D., Vinterbo, S., & Ohno-Machado, L. (2011). Anomaly and signature filtering improve classifier performance for detection of suspicious access to EHRs. In *Proceedings of AMIA Annual Symposium* (Vol. 2011, pp. 723–731).

Koren, Y., Bell, R., & Volinsky, C. (2009). Matrix factorization techniques for recommender systems. *Computer*, *42*(8), 30–37.

Kubat, M., & Matwin, S. (1997). Addressing the curse of imbalanced training sets: one-sided selection. In *Proceedings of the fourteenth international conference on machine learning* (pp. 179–186). San Mateo: Morgan Kaufmann.

Malin, B., Nyemba, S., & Paulett, J. (2011). Learning relational policies from electronic health record access logs. *Journal of Biomedical Informatics*, *44*(2), 333–342.

Menon, A. K., Chitrapura, K. P., Garg, S., Agarwal, D., & Kota, N. (2011). Response prediction using collaborative filtering with hierarchies and side-information. In *KDD'11* (pp. 141–149). New York: ACM.

Murphy, A. H., & Winkler, R. L. (1977). Reliability of subjective probability forecasts of precipitation and temperature. *Journal of the Royal Statistical Society. Series C. Applied Statistics*, *26*(1), 41–47. http://www.jstor.org/stable/2346866.

Office of Technology Assessment, United States Congress (1986) Federal government information technology: electronic record systems and individual privacy, ota-cit-296.

Ornstein, C. (2008). *Fawcett's cancer file breached*. Available online: http://articles.latimes.com/2008/apr/03/local/me-farrah3.

Porter, H. (2010). *Opting out of nhs spine*. Available online: http://www.hhs.gov/ocr/privacy/hipaa/administrative/breachnotificationrule/breachtool.html.

Thai-Nghe, N., Drumond, L., Horváth, T., Nanopoulos, A., & Schmidt-Thieme, L. (2011). Matrix and tensor factorization for predicting student performance. In *CSEDU* (Vol. 1, pp. 69–78).

Wright, A., Soran, C., Jenter, C., Volk, L., Bates, D., & Simon, S. (2010). Physician attitudes toward health information exchange: results of a statewide survey. *Journal of the American Medical Informatics Association*, *17*(1), 66–70.

Yang, S. H., Long, B., Smola, A., Sadagopan, N., Zheng, Z., & Zha, H. (2011). Like like alike: joint friendship and interest propagation in social networks. In *WWW'11* (pp. 537–546).

Zhang, W., Gunter, C., Liebovitz, D., Tian, J., & Malin, B. (2011). Role prediction using electronic medical record system audits. In *Proceedings of AMIA annual symposium* (pp. 858–867).