Journal of Biomedical Informatics 61 (2016) 97–109

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

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

Optimizing annotation resources for natural language de-identification via a game theoretic framework

Muqun Li [a,*], David Carrell [b], John Aberdeen [c], Lynette Hirschman [c], Jacqueline Kirby [d], Bo Li [a], Yevgeniy Vorobeychik [a], Bradley A. Malin [a,e]

[a] Dept. of Electrical Engineering & Computer Science, Vanderbilt University, Nashville, TN, United States
[b] Group Health Research Institute, Seattle, WA, United States
[c] The MITRE Corporation, Bedford, MA, United States
[d] Vanderbilt Institute for Clinical and Translational Research, Vanderbilt University, Nashville, TN, United States
[e] Dept. of Biomedical Informatics, Vanderbilt University, Nashville, TN, United States

ARTICLE INFO

Article history:
Received 17 October 2015
Revised 4 March 2016
Accepted 23 March 2016
Available online 25 March 2016

Keywords:
Electronic medical records
Privacy
Natural language processing
Game theory

ABSTRACT

Objective: Electronic medical records (EMRs) are increasingly repurposed for activities beyond clinical care, such as to support translational research and public policy analysis. To mitigate privacy risks, healthcare organizations (HCOs) aim to remove potentially identifying patient information. A substantial quantity of EMR data is in natural language form and there are concerns that automated tools for detecting identifiers are imperfect and leak information that can be exploited by ill-intentioned data recipients. Thus, HCOs have been encouraged to invest as much effort as possible to find and detect potential identifiers, but such a strategy assumes the recipients are sufficiently incentivized and capable of exploiting leaked identifiers. In practice, such an assumption may not hold true and HCOs may overinvest in de-identification technology. The goal of this study is to design a natural language de-identification framework, rooted in game theory, which enables an HCO to optimize their investments given the expected capabilities of an adversarial recipient.

Methods: We introduce a Stackelberg game to balance risk and utility in natural language de-identification. This game represents a cost-benefit model that enables an HCO with a fixed budget to minimize their investment in the de-identification process. We evaluate this model by assessing the overall payoff to the HCO and the adversary using 2100 clinical notes from Vanderbilt University Medical Center. We simulate several policy alternatives using a range of parameters, including the cost of training a de-identification model and the loss in data utility due to the removal of terms that are not identifiers. In addition, we compare policy options where, when an attacker is fined for misuse, a monetary penalty is paid to the publishing HCO as opposed to a third party (e.g., a federal regulator).

Results: Our results show that when an HCO is forced to exhaust a limited budget (set to $2000 in the study), the precision and recall of the de-identification of the HCO are 0.86 and 0.8, respectively. A game-based approach enables a more refined cost-benefit tradeoff, improving both privacy and utility for the HCO. For example, our investigation shows that it is possible for an HCO to release the data without spending all their budget on de-identification and still deter the attacker, with a precision of 0.77 and a recall of 0.61 for the de-identification. There also exist scenarios in which the model indicates an HCO should not release any data because the risk is too great. In addition, we find that the practice of paying fines back to a HCO (an artifact of suing for breach of contract), as opposed to a third party such as a federal regulator, can induce an elevated level of data sharing risk, where the HCO is incentivized to bait the attacker to elicit compensation.

Conclusions: A game theoretic framework can be applied in leading HCO's to optimized decision making in natural language de-identification investments before sharing EMR data.

* Corresponding author at: 2525 West End Avenue, Suite 1030, Department of Biomedical Informatics, Nashville, TN 37205, United States.
E-mail address: muqun.li@vanderbilt.edu (M. Li).

http://dx.doi.org/10.1016/j.jbi.2016.03.019
1532-0464/

1. Introduction

The past several decades have supported steady growth in the adoption of electronic medical record (EMR) systems [1,2]. While these systems can improve clinical care and efficiency of healthcare operations [3–5], it is increasingly recognized that the data in such resources can be repurposed to enhance various secondary endeavors, such as public health [6,7] and biomedical research [8,9]. To realize such programs on a large scale, it is critical to share EMR data with researchers within and beyond the healthcare organization (HCO) at which it was generated [10]. In certain instances, such as when research is sponsored by the National Institutes of Health, HCOs must have plans for sharing data [11]. There are, however, concerns that the dissemination of such data could infringe upon the privacy rights of the corresponding patients [12,13].

One way to mitigate these concerns, as recommended by the NIH data sharing policy, is to de-identify the data by removing personally identifying information (PII), according to a regulatory standard, such as that specified in the Privacy Rule of the Health Insurance Portability and Accountability Act of 1996, or HIPAA [14]. The de-identification of medical data is relatively straightforward when it is readily apparent where potential identifiers reside (e.g., a column in a database table labeled as "Patient Name"). Yet, a significant portion of EMR data is composed of natural language (e.g., clinical narratives) [15] and current de-identification procedures, whether manual or automated, are unlikely to ever detect all instances of identifiers (i.e., achieve perfect recall) while leaving all instances of non-identifiers in place (i.e., achieve high precision) [16–21]. This implies that no matter how much a HCO invests in the natural language de-identification process, some potentially identifying terms will be leaked or the shared data will not be useful due to redaction of non-identifying terms. This is important to recognize because evidence suggests there are diminishing returns in improving recall and precision as the amount of training data provided to de-identification tools grows.

Given such limitations, and the fact that research budgets (either allocated by the institution internally or obtained through extramural grant support) an HCO needs to determine how much effort should be invested in de-identification. The answer to this question depends, in no small part, on how much shared data is worth – to both the HCO providing the data and the potential recipients (who may exploit it maliciously, such as re-identification of patient data). Therefore, for an HCO to a make rational decision about how much to invest in de-identification, the incentives (and disincentives) of sharing data, as well as the cost-benefit model that incorporates behavior of the anticipated recipients, need to be well-defined.

The goal of our research is to investigate the threat of PII exposure in natural language de-identification as one of (dis)incentives and, subsequently, design a system that minimizes the expenditures of an HCO. Specifically, we model the HCO as a *defender/publisher* who has a limited budget, with a responsibility to protect patient privacy, and the malicious data recipient as a potential *attacker* who attempts to exploit it via re-identification. Under this model, the HCO incurs a cost when performing de-identification (e.g., paying readers to manually redact identifiers or annotate an EMR corpus to train an automated tool) based on which the publisher aims to achieve better protection of the data while retaining its utility. The attacker, by contrast, is incentivized to expose as much sensitive information from the published records as possible, but is bounded in capability (e.g., by a budget of their own) to perform the attack.

We formalize the interaction between the HCO and the ill-intentioned data recipient in a game theoretic framework. In this game, the publisher is a leader, who chooses whether or not to share data, and, if so, how much of their budget to spend on de-identification tasks (with the incentive to minimize spending, so that the surplus may be applied to other activities, such as additional research studies). The attacker, by contrast, is a follower, who aims to discover leaked instances of PII. The publisher may choose not to share the data, for example, if de-identification costs or risks from data sharing outweigh the benefits. The attacker, similarly, may opt out of attacking altogether if the benefits (e.g., from finding and exploiting leaked sensitive information) are not worth the cost of uncovering this information. One important aspect of our framework is that it explicitly models several mechanisms by which an attacker may be deterred. The first is for the publisher to manipulate the data and influence the confidence the attacker has in their claims of identifier discovery. An example of such a strategy is the "hiding in plain sight", or HIPS, approach, whereby all detected instances of identifiers are replaced with fake instances that exhibit a similar semantics (e.g., replacing the name "Rachel" with "Alice", replacing an actual date "4/12/2015" with a randomly generated date "4/25/2015" and replacing a real medical record number "12638920" with a generated medical record number "53267935") [22] which makes it difficult for an attacker to distinguish between fake and real PII. A second deterrence mechanism is to institute data use agreements that penalize the attacker when they commit an exploit and are caught in the act. The model we introduce explicitly represents and reasons over both mechanisms.

This paper provides three primary contributions:

(1) **An adversarial model for natural language de-identification:** The traditional view on natural language de-identification is depicted to the left of Fig. 1. In this view, a publisher considers only the precision and recall of the redaction strategy. The rate of PII discovery tends to grow logarithmically in the amount of training data supplied [17], which means that a publisher would require infinite investment to achieve perfect data protection. However, in the game view, the role of an attacker can be formalized, depicted to the right of Fig. 1, as can the budgets available to both players in the system. In this augmented scenario, both sides engage in a cost-benefit analysis, which explicitly accounts for the interactions between the two agents.

(2) **A Stackelberg formulation of de-identification:** Based on the adversarial model, we introduce a game theoretic approach to solving this problem. This approach is based on a Stackelberg (or leader–follower) game, where the publisher can simulate the capabilities of the adversary before deciding on which strategy to implement (e.g., how much funding to invest in the de-identification process). In doing so, the publisher assumes that the adversary optimizes their strategy and chooses a level of investment in de-identification that maximizes their benefit accounting for an attacker's response.

(3) **Systematic policy evaluation:** We investigate the game under several policy designs for how penalties are paid for violations. We use a dataset of approximately 2100 real clinical notes from Vanderbilt University Medical Center to assess each policy. In doing so, we perform a sensitivity analysis on the decisions made as a function of the costs (e.g., penalties) enforced in the system. We find that there are cases in which the attacker will choose to forgo an attack while the publisher invests only a moderate amount in supporting de-identification. We also show that there are cases when the publisher should choose not to play and not share data.