Journal of Biomedical Informatics 66 (2017) 231–240

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

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

A risk-based framework for biomedical data sharing

Fida K. Dankar [a,*], Radja Badji [b]

[a] College of IT, UAEU, P.O. Box 15551, Al Ain, United Arab Emirates
[b] Sidra Medical and Research Center, P.O. Box 26999, Doha, Qatar

ARTICLE INFO

Article history:
Received 1 June 2016
Revised 15 December 2016
Accepted 19 January 2017
Available online 23 January 2017

Keywords:
Context-aware privacy
Biomedical data
Data disclosure control
Honest broker systems
Institutional review boards

ABSTRACT

The problem of biomedical data sharing is a form of gambling; on one hand it incurs the *risk* of privacy violations and on the other it stands to *profit* from knowledge discovery. In general, the risk of granting data access to a user depends heavily upon the data requested, the purpose for the access, the user requesting the data (user motives) and the security of the user's environment. While traditional *manual* biomedical data sharing processes (based on institutional review boards) are lengthy and demanding, the *automated* ones (known as honest broker systems) disregard the individualities of different requests and offer "one-size-fits-all" solutions to all data requestors. In this manuscript, we propose a conceptual risk-aware data sharing system; the system brings the concept of risk, from all *contextual* information surrounding a data request, into the data disclosure decision module. The decision module, in turn, imposes mitigation measures to counter the calculated risk.

1. Introduction

Multiple government sponsored projects, such as Decode Genetics [1], Korean Reference Genome Project [2], Genome England [3], and recently Qatar National Genome Project [4] have stimulated an outburst in clinical and genomic data stockpiled in biomedical data warehouses. Access to this data offers a unique opportunity to undertake biomedical research and may improve quality of care, reduce healthcare costs and advance personalized treatments. The availability of such data for widespread research activities is dependent on the protection of participants' privacy [5]. In other words, to be able to share participants' information with second parties, the participants' personal information should be protected from being disclosed.

Ethical biomedical data sharing is defined and regulated via social and legal mechanisms (Common rule [6], HIPAA [7], EU data protection directive [8]). Data sharing platforms have to specify technical and manual controls (1) to meet the specified laws and policies and (2) to uphold public expectations of privacy. Several approaches can be employed (separately or jointly) toward this end:

One approach provides data subjects with control over who can access their data, for what purposes and for how long. Users' preferences are usually collected in the form of an informed consent at the time of data collection. However, this approach suffers multiple shortcomings. The current paper-based consent process is static and locks consent information to a single time point (typically during sample collection) [9]. The current process also requires limiting the amount of information conveyed to participants to ensure that their consent is informed, since individuals can only absorb so much information at any one time. Re-contacting participants to obtain additional consents or to provide additional education materials is arduous, time-consuming and expensive. Thus, this approach is not adequate for (multi-purpose) biomedical data warehouses. On the other hand, it was shown in multiple studies that the consent process can bias the participants' pool [10].

Another approach enables data sharing with third parties without consent when the *privacy risk* is low. Research Data requests will be granted, if and only if, privacy risk (i.e. the risk of information disclosure) incurred by such access is acceptable and controlled. Platforms implementing this approach employ Institutional Review Boards to review data sharing applications and perform risk estimation, and/or they apply heuristics/statistical methods (such as de-identification), to protect shared information and lower the privacy risk. However, existing platforms implementing this approach suffer several shortcomings. They are not flexible, they do not reflect all the legal and ethical regulations in the biomedical domain, and, in particular, they cannot impose constraints on data accesses that correctly reflect the privacy risk incurred from data sharing. In [11], the authors define

* Corresponding author at: College of IT, UAEU, P.O. Box 15551, Al Ain, United Arab Emirates.
E-mail addresses: fida.dankar@uaeu.ac.ae (F.K. Dankar), rbadji@sidra.org (R. Badji).

http://dx.doi.org/10.1016/j.jbi.2017.01.012
1532-0464/

the privacy risk of a data request as a predictive function of the expected value of damage:

$$\text{Quantified risk} = (\text{prob of damage}) \times (\text{value of damage})$$

Such damage can be caused by a number of circumstantial factors that are specific to the data sharing episode such as the sensitivity of the requested data and the trustworthiness of the user [12–15]. Thus, to express privacy risk, one has to be able to define and measure all the necessary components affecting this estimation.

In this paper we propose a risk-aware information disclosure model for biomedical data. Our model evaluates the risk posed by a data request using all *contextual information surrounding the request* and feeds it into an access control decision module. In turn, the decision module imposes mitigation measures to counter the posed risk. The concept of risk-aware information disclosure was introduced in [15] where the impediments of existing data sharing approaches was discussed along with initial ideas on risk-aware access without articulation of a precise and complete model.

2. Related research and contributions

As indicated previously, the study of (non-consented) identifiable biomedical data requires approval from an Institutional Review Board, IRB (also known as Research Ethics Boards in certain countries). IRB procedures are extensive and can obstruct timely research and discoveries [16–18]. Studies on platforms that rely on IRB for all data accesses reveal unsatisfied users. The application process is strenuous and approvals take a long time often delaying project initiation significantly [18,19].

To reduce complications, many countries enacted regulations that permit data sharing without IRB approval when data is believed to be anonymous. Following this regulation, many platforms implement an automated Honest Broker System (HBS) to provide de-identified data to individual investigators [20–22]. In such cases, the original database goes through a de-identification algorithm and the result is stored in a separate database. The de-identified database is maintained by the HBS and made available to individual investigators upon request. Although these platforms reduce significantly the time to acquire data, they offer one de-identification for all data requests. If the employed de-identification is stringent the utility of the data will be affected, and investigators might prefer to apply to the IRB. If the de-identification is light, then the data holder's concern about patient privacy would persist, leading to conservative disclosure decisions (withholding data from unknown or non-trusted investigators).

To illustrate the aforementioned problems, consider the following two scenarios:

> *An investigator from a newly established research institute requests access to the records of HIV infected subjects, the purpose is to perform a study on gene expression in HIV infection.*
> *An investigator from a well-established research institute with recognized privacy and security practices requests access to the records of flu-infected people to perform a study on flu vaccine effectiveness.*

A responsible data-sharing mechanism would impose more mitigation measures on the first request due to the high sensitivity of the data, and its potential for injury. Such mitigations could manifest as reduction in the granularity of the data (de-identification) and/or as restrictions on when and how a user can access the data.

2.1. Related research

2.1.1. Risk modeling

Risk-aware access control received growing attention in the past few years. Most research in this domain gave little attention to quantifying the risk posed by a data request and focused rather on designing models for policy enforcement [14,23–25]. Moreover, these policy enforcement models have fixed access rules and do not allow legal entities (such as the IRB) to override system rules, in other words, they do not allow exceptions.

Nonetheless, there are some efforts related to the quantification of privacy, among which we cite the following:

- In an effort to quantify data privacy, Westin [26] uses social science to understand privacy beliefs of different data providers, and Ngoc et al. [27] uses information theory to calculate information leakage due to a data release, however, these approaches cannot be generalized as they concentrate on specific problems.
- Adams [28] attempts to model users' perceptions of privacy in multimedia environments. He identified three factors that determine users' perceptions of privacy: information sensitivity (user's perception of the sensitivity of the released information), information receiver (the level of trust the user has in the information recipient(s)) and information usage (costs and benefits of the perceived usages). Lederer [29] uses Adams' model as a framework for conceptualizing privacy in ubiquitous computing environments in addition to the Lessig model [30] for conceptualizing the influence of societal forces on the understanding of privacy. While very helpful in understanding the different dimensions of privacy risk, these efforts concentrate on privacy quantification from the participant perspective rather than the data holder.
- Barker et al. [31] introduces a 4 dimensional model for privacy: purpose (data uses), visibility (who will access the data), granularity (data specificity) and retention (time data is kept in storage). Barker's et al. model was later used by Banerjee et al. [32] to quantify privacy violations.
- In multiple consecutive studies [12,33], El Emam et al. defined three criteria that contribute to data identifiability, these are users' motives, the sensitivity of the requested data, and the security controls employed by the data requestor. The authors state that, according to their long experience in private data sharing [12,34–36], these are the main criteria used (informally) by data custodians.

2.1.2. Data protection

Data protection methods (or mitigation measures) can be divided into two broad categories: process driven and data driven [15]. In process driven mechanisms, the dataset is held by a trusted server, users query the data through the server and privacy is built into the algorithms that access the data. Differential privacy is the most popular process-driven privacy model [37]. It perturbs queries' output in a random but controlled manner. Differential privacy requires that the answer to any query be "probabilistically indistinguishable" with or without a particular row in the database. Precisely, given two databases that differ in exactly one row, a differentially private algorithm will provide randomized outputs that follow almost identical probability distributions on both databases. Differential privacy provides strong proofs for privacy, but often leads to low utility particularly in studies that rely on rare events (such as association studies that look at rare genetic events) [38]. In such events, it could lead to strange data representations and erroneous associations [39].

Data driven approaches suppress or modify the variables that could allow an attacker to know precisely the owner of a particular