PSST... Privacy, Safety, Security, and Trust in Health Information Websites

Hamman W. Samuel
Dept. of Computing Science
University of Alberta
Edmonton, Alberta, Canada
hwsamuel@cs.ualberta.ca

Osmar R. Zaïane
Dept. of Computing Science
University of Alberta
Edmonton, Alberta, Canada
zaiane@cs.ualberta.ca

Abstract— **Various newsworthy incidents typically include breaches of security, invasion of privacy, and harm caused by false information. In the e-health domain, there has been a lot of focus on ethical issues when dealing with electronic health records (EHRs) and patient medical records (PMRs). However, equally important are the myriad of health information websites that are being used to formally or informally get medical advice online. This study surveys related work on three popular and pertinent issues in health information websites: privacy, security, and trust. Our contributions include a succinct survey of different categories of popular health information websites (WebMD.com, MayoClinic.com, KidsHealth.org, PatientsLikeMe.com) to gauge existing methods for handling these issues. Moreover, an agenda is proposed for understanding the three issues orthogonally via access control. Other outcomes of the study include recommendations for open problems identified in health websites, including the need for fine-grained privacy, security and trust controls.**

I. Introduction

There is a lot of focus on addressing ethical issues arising with Electronic Health Records (EHRs) and Patient Medical Records (PMRs). However, an equally important area is health information websites and health social networking services that are gaining more prominence and popularity. Health websites with communities such as WebMD.com and PatientsLikeMe.com have features for users to register, which leads to storage of user information. This data can be considered sensitive and can make the user susceptible to privacy attacks. Considering the case of the AOL search logs incident, user data from other sites can be used to re-identify other anonymized records via various linkage attack models[8].

Furthermore, even in purely informational health websites with no user communities, there is the risk of inaccurate health information. For instance, in the United States, websites like ApricotsFromGod.info continue to operate and publicize cures for cancer, without the approval of the Food and Drug Administration (FDA)[2]. ApricotsFromGod.info had 1,532 visitors in the past year according to statistics from Compete.com, who were exposed to the risk of taking incorrect and untested health advice.

Health information is defined by the Internet Health Coalition (IHC) as follows: 'Health information includes information for staying well, preventing and managing disease, and making other decisions related to health and health care'[4]. There are various kinds of websites, such as blogs, wikis, forums, social networks, and so on. Health information websites have varying and unique features and ethics requirements that fit different classifications. In this study, the prominent issues in different categories of health information websites are investigated. Popular health websites within different categories are used to study these ethics requirements. Solutions are proposed to the identified problems, while keeping focus on how much technology can realistically achieve.

II. Ethics Requirements of Health Information Websites

Popular websites from within the two broader categorizations are used to study ethical issues that arise. The following classifications are based on our previous work on creating a taxonomy of health information websites[20]. This taxonomy can also be used to define two broad classifications of health websites: **article-based** and **community-based**. Content on article-based health websites is Owner-Engineered Content (OEC), and there is none of User-Generated Content (UGC) present. In contrast, a community-based health website can have a mix of both content types. The distinguishing factor is users' ability to contribute to the content base of the website.

A. Article-Based Health Information Websites

MayoClinic.com and KidsHealth.org are two popular websites, with traffic of over eight million and two million last year respectively according to Compete.com. MayoClinic.com features general-topic articles written by the Mayo Clinic staff on various health topics, diseases, symptoms, drugs, tests, and so on. The website also features blogs written by experts. KidsHealth.org targets parents, children, and teens. The website predominantly contains reviewed articles on general topics. Within these types of websites, privacy is seemingly not an issue because no personal data is being collected.

In terms of trustworthiness of content, MayoClinic.com provides details and profiles about the authors of their expert blogs, while most articles are labeled with 'Mayo Clinic Staff' as the author. MayoClinic.com also displays the HONCode certification logo. KidsHealth.org gives names and profiles of reviewers of each article at its end. The website has an expert review board as well as medical editors[15], [11]. Since possibly no medical data of guest users is collected, privacy and security concerns and risks are toned down or non-existent.

However, the issue of trust still arises. Trust in the context of article-based health websites can also be defined in terms of the truthfulness of the content. For example, different individuals have varied reactions and allergies that the article may not have highlighted. Also, there is the likelihood that the article contains outdated, incomplete, or even inaccurate information. Studies have also shown that not all websites that display certification logos may be in full compliance with the guidelines of these certification bodies[10]. Guaranteeing trustworthiness of information still remains an open problem.

B. *Community-Based Health Information Websites*

PatientsLikeMe.com is a social network with a vibrant community where patients can share advice, stories, and treatments with others going through their experiences. There are two privacy settings provided: 'visible' and 'public'. PatientsLikeMe.com also incorporates a visual star-based user rating in each user's profile. These ratings seem to provide some indication of trustworthiness of the user[18].

WebMD.com doubles up as both an article-based and a community-based health website. Discussions within communities allow registered users to write opinions about any topics they are interested in, and also contribute to other users' discussions. In its privacy policy, WebMD.com provides comprehensive details about how it ensures security by authentication, authorization, backups, and audit trails. WebMD.com's community is more forum- or bulletin board-based than a social network. Hence, there are no specific privacy settings available for the user[22].

For SNS, since users can register and store their personal information, issues of privacy can come into play. Even if users do not store sensitive medical data upon registration, part of their personal information is being exposed, such as name, email, etc. It is possible to use this partial data in re-identification through linkage attacks[8]. Relational information is equally important because security in SNS involves the additional task of considering the reach of a user's network[9], [12]. However, for communities in PatientsLikeMe.com and WebMD.com, 'friend' relationships are not defined, so a pure social graph is not present[22], [18].

C. *Cross-cutting Ethical Issues*

When intersecting the ethics requirements with potential issues, the following needs and improvements can be identified within health websites.

1) Better Privacy: An open problem we highlight in health websites is the lack of fine-grained privacy controls. Generally, parts of a profile may be more sensitive than others. While most websites look at privacy control in terms of a block of profile fields, users may wish to keep some fields more private than others[14]. For instance, a user may not want all their friends to see their year of birth, but still want to announce their birthday dates. In health websites, a similar situation can arise. Users may be asked to enter their health insurance information, or hospitalizations, such as on PatientsLikeMe.com. The user may wish to share this with only select close family and friends, not generally with registered users. An equally important aspect of privacy controls is that too many privacy settings would be overwhelming for the user[14]. Fang and LeFevre propose a privacy wizard that can learn from the user's basic privacy preferences[7]. However, this method may not be as applicable when sensitive data is involved, because misclassification errors could have more serious results.

2) Credibility of Health Information: Because health websites involve using health advice, trust or credibility is an essential ethical requirement. Trusting a website also entails believing that it presents factual and up-to-date health information, and trust is ultimately related to the source of the information[17]. Studies on identifying trust metrics for health websites have found that user interface elements and aesthetic qualities add to credibility[19], [6]. However, a well-designed or attractive interface may give a false sense of security, because it says nothing about the quality of health content. Moreover, certification logos can also give a false sense of credibility. A recent study found that only 66% of the websites that displayed the HONcode logo were actually in full compliance[10]. Also, star-based reputation systems are inadequate in fully measuring trustworthiness of the user and any content they may generate. This is because if users do not participate or provide ratings, their reputation scores would be low, which does not necessarily reflect their credibility, but rather their activeness.

3) The Role of Technology: Security mechanisms such as authorization and authentication ensure privacy to an extent. Also, rating systems and logos can give an idea of the credibility of a website. However, technological solutions to ethical issues are limited by the extent to which the user can effectively use them. The Internet is moving towards a virtual representation of the real world. However, there are no physical restrictions such as walls or locks and all security and credibility measures are abstract and circumventable. In this situation, technology can at best serve as an enabler, but humans are ultimately responsible for ensuring ethical requirements.

III. Proposed Solutions via Access Control

When looking at different breaches, of privacy, security, or trust, certain points stand out: invasion of privacy is caused by unauthorized access to personal data; a security breach usually involves by-passing authentication to access data; breaches of trust begin with technical flaws or sinister motives that affect data[1]. In health information websites, especially Health Social Networking Services (HSNS), an authorized and authenticated user may share health advice that is incorrect, incomplete, out of date, or inapplicable to other users, leading to a breach of trust. The common theme in these breaches is controlling access to data. Access control is a generic term for authorization and authentication. It involves allowing or disallowing a user from performing a set of operations on some given resource[5]. Access control is usually performed using security policies agreed on by security administrators or users[3]. Access control decisions are a combination of subjects/users, objects, and permissions, eventually leading to a user being granted or denied permission to access data in some context[21]. Consequently, authenticated and authorized users can be seen as trustworthy because the access control decision had been made based on the information owner's allowances.

A. Access Control Model

More formally, access control can be expressed in terms of the following basic components, based on [16] and [3].

- A set of subjects, S, that authorization needs to be given to, such as users
- A set of objects, O, that need to be protected, for example, data
- A set of permissions, P, assigned to subjects on objects
- A set of actions, A, that can be performed in the context of the permissions given

Given the above components, access control involves defining $P_i(S_i, O_i), \forall S_i \in S, O_i \in O, P_i \in P$. In the context of health information websites, P defines authorizations given to users to access certain pages or data. For instance, a user may not be allowed to see another user's profile because of privacy settings. In addition, given P_i, S_i could perform one of the following fundamental operations on O_i[13].

- *Create, C*: S_i can be allowed to create new instances of O, $C = \{1 \mid 0\}$, where 1 implies allowing and 0 implies disallowing
- *Read, R*: S_i can be allowed to view O_i, $R = \{1 \mid 0\}$
- *Update, U*: S_i can be allowed to change O_i, $U = \{1 \mid 0\}$
- *Delete, D*: S_i can be allowed to remove O_i, $D = \{1 \mid 0\}$

Now A can be defined as $A_i = C_i \cup R_i \cup U_i \cup D_i$. Consequently, an access control decision can be modeled as a quadruple, $AC(S, O, P, A)$.

B. Orthogonal View of Privacy, Security, and Trust

We also present an abstracted orthogonal description of access control incorporated with notions of privacy, security, and trust using graph notation. Let $G = (V, E)$ be a directed graph, where G is the system, V are nodes representing entities containing information, and E are edges representing access attempts. We refer to entities as an abstraction of users, user profiles, content, comments, and the like. A typical attempt at accessing information would be represented as (i, j), where i is the entity requesting the information, and j is the node source of the information, given $i, j \in V$. We can define privacy, Pr as a *label*, or more formally a weight, on (i, j), such that $Pr(j, i) = \{1 \mid 0\}$.

Here, 1 represents allowing and 0 represents denying access to an action. For simplicity, we abstract the fundamental operations as one label, and note that a more comprehensive label can simply be defined as a 4-tuple notation $(\{1 \mid 0\}, \{1 \mid 0\}, \{1 \mid 0\}, \{1 \mid 0\})$ corresponding to (C, R, U, D) if required. $Pr(j, i)$ is a configuration of the privacy of entity j with respect to access requests from entity i. The proposed notation defines an access-specific privacy of j, which is represented on the directed graph by the edge (i, j) because i is attempting access to information about j. A base privacy definition also needs to be established. This is simply $Pr(i, i) = 1$, so that an entity always has self-access. Privacy can consequently be defined as a piecewise function.

$$Pr(j,i) = \begin{cases} \{1 \mid 0\} & : i \neq j \\ 1 & : i = j \end{cases}$$

In addition, default privacy is defined as $Pr(j, i) = 0$ for the situation where no edges or labels are present. Consequently, security can be defined as the access control decision after reading the label for (i, j). Trust for i can be defined in terms of a simplistic score using inbound edges to i, E_i, such that $Tr(i) = n(E_i) + \sum Pr(k, i), \forall k \in V, k \neq i$. For a health website, this definition of trust implies that a user who has been allowed access by many other users is probably more trustworthy than someone who has not been. This follows the definition of trust as the 'willingness to be vulnerable'. Figure 1 gives an example of the proposed abstraction model. In the example, entity A is sharing data with entity B. However, entities B and C have disallowed access to their data for all other entities. By the definition of trust in terms of access, A is the most trustworthy.

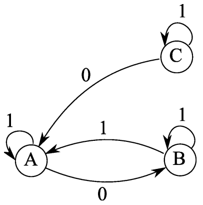

Fig. 1. Example Graph Notation Representation of Privacy, Security and Trust

Privacy of A *wrt* B = (B, A) = 1
Privacy of A *wrt* C = (C, A) = 0
Privacy of B *wrt* A = (A, B) = 0
Privacy of B *wrt* C = (C, B) = ∅ = 0
Privacy of C *wrt* B = (A, C) = ∅ = 0
Privacy of C *wrt* C = (B, C) = ∅ = 0
Trust score of A = 2 + 1 = 3
Trust score of B = 1 + 0 = 1
Trust score of C = 0 + 0 = 0

C. Fine-Grain Privacy

The granularity of privacy controls can be viewed in parallel with lock granularity in database management systems. Our interest is at the record level and going further fine-grained. Typical privacy settings focus on record entities or logical blocks of data. This is more coarse-grained than the ethical requirements of health websites. Field-level granularity requires that each field have its own privacy configuration per record. In terms of the orthogonal model, fine-grain privacy can be represented by n-tuples of labels, instead of having just one label. So instead of $Pr(j, i) = \{1 \mid 0\}$, we could have $Pr(j, i) = (\overline{f_1}, \overline{f_2}, ..., \overline{f_n})$, where $\overline{f_i}$ is a privacy setting for a field, f_i. A simplistic example of this is shown in Figure 2. It is normal to have the actual implementation of hiding or showing data happening at the application layer. Our method is aimed at storing users' field-level privacy configurations. Implementation-level details are in line with this conceptual explanation. Let $\tau(\widehat{f}, f_1, f_2, ..., f_n)$ define a table within a health website database, where f_i represent fields within τ, and $\widehat{f}$ represents the primary key. Also, let $v(\widehat{u}, ...)$ describe the users' table, where $\widehat{u}$ is the primary identifying key for each user. In addition, let $(\underline{f_1}, \underline{f_2}, ..., \underline{f_n})_i$ represent the ith record in τ. We then define a new table for storing privacy settings for each field in τ, $\rho(\widehat{f}, \widehat{u}, \overline{f_1}, \overline{f_2}, ..., \overline{f_n})$, where $\overline{f_i}$ is a field that represents the privacy setting for the ith record and specifically the field f_i in τ, i.e cell $\underline{f_i}$. Simplistically, $\overline{f_i}$ can be a boolean value that represents showing or hiding the data associated with the field.

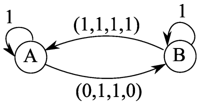

Fig. 2. Extended Example Representation for Fine-Grain Privacy

D. Trust Metrics

In PatientsLikeMe.com, privacy and trust are distinct from the orthogonal representation of privacy and trust as Pr and Tr respectively. Privacy in this case is not based on interactions, but is related to the user alone, and not the requester. In addition, trust is not being measured by access rights or authorization in PatientsLikeMe.com, but by keeping external scores or ratings. Similarly, article-based websites do not conform to the earlier definition of trust as $Tr(i)$, because there are no interactions to measure trust in that way. To address these issues, we first extend the orthogonal description of privacy and trust by extending the label concept to include trust ratings in addition to privacy. An access attempt can be labeled with fine-grain privacy settings, and also contain trust ratings and other trust metrics. A simplified representation is presented in Figure 3. Privacy and trust are defined as a 2-tuple weight on the edges.

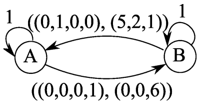

Fig. 3. Extended Representation for Fine-Grain Privacy and Trust Metrics

IV. Conclusion

Breaches of security, invasion of privacy, and harm caused by false information make for newsworthy incidents. In spite of these problems, more and more health information websites are being used to get medical advice online. We carried out a survey of existing health websites to identify ethical requirements such as privacy, security and credibility within different categorizations. Finally, we commented on the role of technology in proposing solutions to reduce risk and potential harm, and proposed solutions to address cross-cutting ethical requirements.

References

[1] Alessandro Acquisti, Rahul Telang, and Allan Friedman. Is There a Cost to Privacy Breaches? An Event Study. In *International Conference on Intelligent Systems*, pages 26–33, 2006.

[2] James Anderson and Kenneth Goodman. *Ethics and Information Technology: A Case-Based Approach to a Health Care System in Transition*. Springer New York, 2002.

[3] Elisa Bertino, Barbara Catania, Elena Ferrari, and Paolo Perlasca. A Logical Framework for Reasoning about Access Control Models. *ACM Transactions on Information Systems Security*, 6:71–127, 2003.

[4] Internet Health Coalition. E-Health Code of Ethics. `http://www.ihealthcoalition.org/ehealth-code-of-ethics/`. Accessed June 23, 2011.

[5] Apache Documentation. Authentication, Authorization, and Access Control. `http://httpd.apache.org/docs/1.3/howto/auth.html`. Newsletter, Accessed June 23, 2011.

[6] Silvana Faja and Adriatik Likcani. E-Health: An Exploratory Study of Trust Building Elements in Behavioral Health Web Sites. *Journal of Information Science and Technology*, 3(1):9–22, 2006.

[7] Lujun Fang and Kristen LeFevre. Privacy Wizards for Social Networking Sites. In *International Conference on World Wide Web*, pages 351–360, 2010.

[8] Benjamin Fung, Ke Wang, Rui Chen, and Philip Yu. Privacy-Preserving Data Publishing: A Survey of Recent Developments. *ACM Computing Surveys*, 42:1–53, 2010.

[9] Seda Gürses and Bettina Berendt. *The Social Web and Privacy: Practices, Reciprocity and Conflict Detection in Social Networks*. Chapman and Hall, CRC Press, Florida, 2010. Privacy-Aware Knowledge Discovery: Novel Applications and New Techniques.

[10] Yi Hong, Timothy Patrick, and Rick Gillis. Protection of Patient's Privacy and Data Security in E-Health Services. In *International Conference on BioMedical Engineering and Informatics*, pages 643–647, 2008.

[11] KidsHealth. Privacy Policy. `http://kidshealth.org/parent/kh_misc/policy.html#cat20186`. Accessed June 23, 2011.

[12] Alexander Korth, Stephan Baumann, and Andreas Nürnberger. An Interdisciplinary Problem Taxonomy for User Privacy in Social Networking Services. In *Workshop on Privacy for a Networked World*, 2011.

[13] James Martin. *Managing the Data Base Environment*. Prentice Hall PTR, Upper Saddle River, 1983.

[14] Michael Maximilien, Tyrone Grandison, Tony Sun, Dwayne Richardson, Sherry Guo, and Kun Liu. Privacy-as-a-Service: Models, Algorithms, and Results on the Facebook Platform. In *Web 2.0 Security and Privacy Workshop*, 2009.

[15] MayoClinic. Privacy Policy. `http://www.mayoclinic.com/health/privacy-policy/AM00005`, 2010. Accessed June 23, 2011.

[16] Mark Miller, Ka-Ping Yee, and Jonathan Shapiro. Capability Myths Demolished. Technical report, 2003.

[17] Health on the Net. Health on the Net Code of Conduct (HONCode). `http://www.hon.ch/HONcode/Webmasters/index.html`. Accessed June 8, 2011.

[18] PatientsLikeMe. Privacy Policy. `http://www.patientslikeme.com/about/privacy`. Accessed June 21, 2011.

[19] Jan Rosenvinge, Stein Laugerud, and Per Hjortdahl. Trust in Health Websites: A Survey among Norwegian Internet Users. *Journal of Telemedicine and Telecare*, 2003.

[20] Hamman Samuel and Osmar Zaïane. HCMS: Conceptual Description of a Health Content Management System. In *3rd Workshop on Software Engineering in Health Care*, pages 17–23, 2011.

[21] Waleed Smari, Jian Zhu, and Patrice Clemente. Trust and Privacy in Attribute Based Access Control for Collaboration Environments. In *International Conference on Information Integration and Web-based Applications and Services*, pages 49–55, 2009.

[22] WebMD. Privacy Policy. `http://www.webmd.com/about-webmd-policies/about-privacy-policy?ss=ftr`, 2011. Accessed June 21, 2011.