New Privacy Threats in Healthcare Informatics: When Medical Records Join the Web

Fengjun Li[†], Jake Y. Chen[‡], Xukai Zou[‡], Peng Liu[†]
[†] College of IST, The Pennsylvania State University, University Park, PA, USA
[‡] Department of Computer and Information Science, IUPUI, Indianapolis, IN, USA
{fli,pliu}@ist.psu.edu, jakechen@iupui.edu, xkzou@cs.iupui.edu

ABSTRACT

In this paper, we study how patient privacy could be compromised from electronic health records (EHRs), especially with the help of today's information technologies. Current research on privacy protection is centralized around EHR: protecting patient information from being abused by authorized users or being accessed by unauthorized users. Limited efforts have been devoted to studying the attacks performed by manipulating information from external sources, or by joining information from multiple sources. Particularly, we show that (1) healthcare information could be collected by associating and aggregating information across multiple online sources including social networks, public records and search engines. Through attribution, inference and aggregation attacks, user identity and privacy are very vulnerable. (2) People are highly identifiable even when the attacker only possess inaccurate information. With real-world case study and experiments, we show that such attacks are valid and threatening. We claim that too much information has been made available electronic and available online that people are very vulnerable without effective privacy protection.

General Terms

Security

Keywords

Privacy, Healthcare informatics, EHR, social networks

1. INTRODUCTION

In recent years, with the development of healthcare informatics, a large amount of medical/healthcare records have been digitalized (in EHRs), for example, 43.9% of the US medical offices have adopted full or partial EHR systems by 2009 [7]. Since medical records are considered to be extremely sensitive, people start to concern on their privacy with digitalized healthcare data. Security and privacy becomes an important and popular topic in healthcare informatics research. Existing research on protecting user privacy in healthcare information systems could be summarized into three categories: (1) Defending against internal abuse of electronic health data, e.g. hospital personnel with authorization to access patients' records disclosing some of the private information for non-medical purposes. (2) Defending against unauthorized access to electronic health data, e.g. attackers hacking into hospital's databases or eavesdropping over the network communications. (3) Defending against re-identification attacks against published electronic health records, e.g. adversaries with access to de-identified healthcare data that are published for research purposes discovering the identities of record owners from a set of unprotected quasi-identifiers.

Meanwhile, as the Web gains its popularity and touches many aspects of our daily life, it becomes the largest open-access source of personal information. First, large amount of public records have been made accessible online, including phone books, voter registration, birth/death records, etc. Although some of them enforce certain restrictions to defend against abusers, it is still relatively easy or inexpensive to crawl/download such records. Second, more recently, online social network sites such as Facebook and MySpace have emerged to successfully attract a huge number of users, who willingly put their personal information to online social network sites to share with people. Unfortunately, with the new sophistication of information retrieval techniques and the advancement of searching techniques in search engines, it becomes unexpectedly easy to conduct Web-scale extraction of users' personal information that is readily available in various online social networks (e.g., [1, 8, 13, 3, 4]). As a result, malicious or curious adversaries could easily take advantage of these techniques to collect others' private information, which is readily available from online public records or various social networks.

In this way, the attackers possess powerful weapons and rich knowledge, which are somehow provided by the victims themselves, and are truly beyond the assumptions in the research literature. In this paper, we ask the question: "*when an attacker possesses a small amount of (possibly inaccurate) information from healthcare-related sources, and associate such information with publicly-accessible information from online sources, how likely the attacker would be able to discover the identity of the targeted patient, and what the potential privacy risks are.*"

To take a first step in answering this broad question, we study: (1) how user information from multiple online sources could be associated and utilized to compromise user privacy;

(2) how user identity could be identified by comparing approximate information with public databases.

2. ATTACKS ON HEALTHCARE RECORDS

With the broad adoption of electronic health records, security and privacy becomes extremely critical. Current researches on protecting patient privacy are centralized around the protection of EHRs by protecting patient information from being: (1) abused by authorized users; (2) accessed by unauthorized parties; or (3) re-identified from healthcare data published for research purposes.

To protect health care related information, regulations for disclosure are set and protected by law [2]. However, healthcare related personnel may violates privacy rules by disclosing or stealing private healthcare records for unauthorized usages, as depicted in [16]. This is a typical abuse/infraction with authorized data access. More often, the attackers do not have authorization for data access. They either eavesdrop or wiretap private information in transit or penetrate into EHR systems to get control of valuable health data. However, such types of attacks are often underestimated [18]. We believe such underestimation is partially from a fundamental misunderstanding that information revealed by carelessness or misuse is only one piece of the big picture and will not cause severe privacy disclosure. In later this paper, we will elaborate the severeness of such type of attacks in current information-rich context with an intuitive example.

Recently, there has been an increasing demand to publish the immense volume of EHRs for secondary purposes, such as research, government management, payment, and other marketing usages [14]. A typical EHR consists of a set of identifier attributes (e.g. *name*, *SSN*), quasi-identifier attributes (e.g. *gender*, *zipcode*), and sensitive attributes (e.g. *diseases*). Since privacy of record owners becomes a major concern, EHRs need to be de-identified [6] or anonymized [15] before data publishing. However, even with de-identified or anonymized data, sensitive attributes that pertains to an individual may be learned from other non-sensitive attributes in combination with external knowledge (e.g. voter registration list, phone books, etc.). The risks of such re-identification attacks have been intensively studied, which shows that the amounts and types of an attacker's external knowledge play an important role in reasoning about privacy in data publishing [11, 9, 12, 5]. However, it is not easy if not impossible for a data publisher to know upfront what external knowledge the attackers possess. Therefore, current research on privacy-preserving data publishing studies the problem from a theoretical perspective by making assumptions on attacker's background knowledge, quantifying external knowledge regardless of its content, and sanitizing the data to ensure the amount of disclosure is below a specified threshold [12, 5]. As a result, such protection, on one hand, does not take into account that large amount of external knowledge are accessible to the adversaries from various online sources (e.g. social networks), on the other hand, it might greatly distort the data and its secondary usages. Therefore, I believe it is of great importance to investigate the types and amounts of external knowledge that a powerful attacker possesses or infers from the immense volume of electronic data from *multiple online resources*. It not only provides evidence for efficient and optimal data sanitization, but also raises public concerns and awareness on the severeness of privacy threats and calls for effective protection.

3. ATTACKS FROM EXTERNAL SOURCES

Recently, online social networks are becoming extremely popular. Participants often voluntarily disclose personal information with surprising details. For example, *LinkedIn* users list their educational and working experiences to seek for potential career opportunities, and *MedHelp* users share details about their life and medical experiences expecting suggestions from others. A fundamental misunderstanding is that it is unlikely to link information of the same individual from different online resources. Unfortunately, with the sophistication of searching and information retrieval techniques, it is feasible for an attacker to *aggregate* personal information of a target user on different online resources, by associating unprotected but identifiable or semi-identifiable attributes (e.g. identical account names or email address of a careless user) [10]. Meanwhile, with governmental and industrial efforts, a large amount of public records have been digitalized and made available online. Most of them are indexed by commercial search engines, while others require a minimum subscription fee for full access. Adversaries could easily access and utilize such information to compromise others' privacy. Especially, it is possible to aggregate and associate information from multiple (possibly medical-related) external sources to identify patients from their poorly-anonymized data and reconstruct their complete profiles including identifiers and quasi-identifiers, as well as sensitive medical information.

Figure 1 demonstrates an example from a real-world case study: "Jean" (whose full name has been discovered but removed here for privacy protection) has type II diabetes, so she actively participates in two medicare social networks, *MedHelp* (www.medhelp.org) and *MP and Th1 Discussion Forum* (www.curemyth1.org). Her profile in *MP and Th1*, as shown in Figure 1 (1), contains birthdate, occupation, location, email addresses, and a text field about her interests on medical information. Her profile in *MedHelp*, as shown in Figure 1 (2), includes gender, age, location, and a text, from which we can learn astonishing details about Jean's medical conditions and history, e.g. *Diabetes II*, and *Ac1=5*, etc. More private attributes of Jean (e.g. times of doctor visit or diagnoses, prescription and medication) could be extracted from her postings on the two sites, respectively.

As shown in Figure 1, we compare all the attributes from both profiles: (1) Jean used identical (and relatively unique) username on both sites; (2) both profiles show Jean's current location - a small town with approximately 15K population; (3) birthdate shown in Profile 1 is consistent with the age shown in Profile 2; (4) Profile 1 shows "my husband" that indicates the owner is a female, which is consistent with the gender shown in Profile 2; and (5) both profiles show the same disease and symptoms. With all the evidences, we are able to link the two profiles at a certain confidence level, and associate the attributes from both profiles to the same individual. Further more, with the email address provided in profile 1, we are able to get profile 4 through Web search engines (note that email addresses are always considered as identifiers). Profile 4 includes a phone number (later it turns out to be a cell phone number) and a P.O. Box address, which also shows the same city as in Profiles 1 and 2. With the phone number from profile 4, we further discovered Profile 3, which is a job-related page containing Jean's cell and home phone numbers. Profiles 3 and 4 both contains the full name of "Jean", and we have a good hint on her occupation.

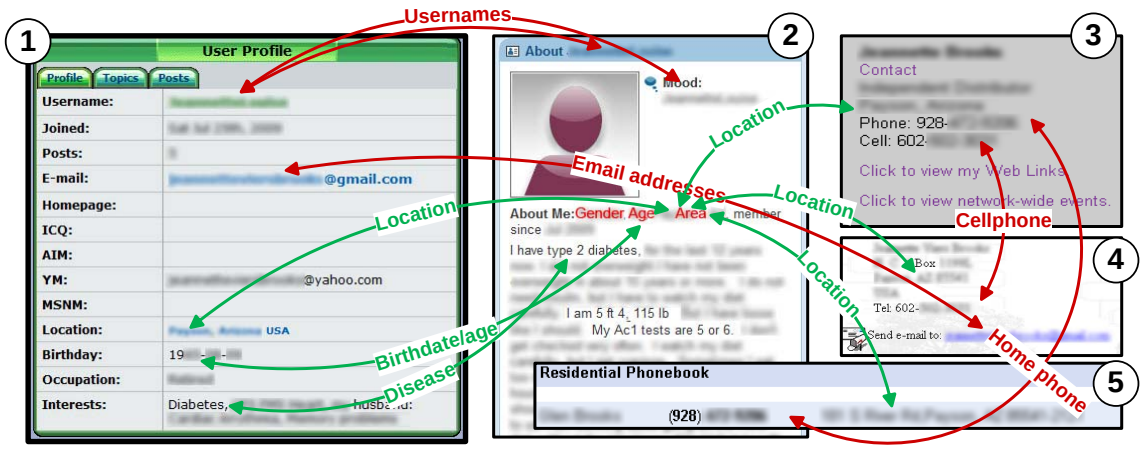

Figure 1: A real-world example of cross-site information aggregation.

Finally, with the home phone number, we are able to locate Jean's record in the residential phonebook, which shows her husband's name and their full home address. On the other hand, even without Profiles 3, 4 and 5, an attacker could also utilize public records to get more information about Jean: with the attribute set {gender, birthday, location}, Jean's identity (e.g. full name, address, and phone number) is recoverable from public birth records, voters registration records or online phone books.

By associating five profiles, we have collected Jean's full name, date of birth, husband's name, home address, home phone number, cell phone number, two email addresses, occupation, medical information including lab test results. With her full name, more information about Jean is subsequently discovered from various social networks. Finally, when Jean's hospital publishes de-identified patient records to support medical research, the attacker with external knowledge obtained from above process is highly likely to re-identify Jean's record.

The example reveals a serious privacy issue in both social networks and healthcare informatics. The entire process includes three steps: *attribution*, *inference*, and *aggregation* attacks. In attribution, identifiable, semi-identifiable or sensitive attributes are learned/extracted from various sources over the web. Particularly, three types of online resources are considered in the example: (1) public-accessible online databases: voters registration records, phone books, birth and death records, (2) online social network sites with explicit identifiable attributes (e.g. *LinkedIn*, *Facebook*, etc.) as well as specified healthcare-related social networks (e.g. *MedHelp*); and (3) commercial search engines, which index a good portion of the web. In inference, more attributes are further discovered from social activities and relationships through statistical learning or logical reasoning. In aggregation, records retrieved from different sources that potentially pertain to the same individual are linked under strong or weak evidences, in which strong evidences include matching identifiers or quasi-identifiers, and weak evidences are similarities identified from a statistical perspective. As we have shown in the example, the attacks are very valid and do not require excessive resources or techniques. Therefore, people are very vulnerable under such attacks, if they do not careful protect their online identities. A powerful privacy protection tool is expected to defend against such attacks.

4. ATTACKS WITH APPROXIMATE INFORMATION

Besides privacy attacks against digitalized medical records and healthcare information systems, adversaries also seek to obtain valuable information with non-technical kind of intrusions such as insider incidents or social engineering. With a vague definition, insider incidents often involve abuses such as inside personnel accidental leaking or stealing information, using pirated software, or accessing questionable webpages. Social engineering relies on people's unawareness of valuable information and carelessness in protection and becomes one of the major attacks towards user privacy. However, in most cases, information obtained from non-digital channels are not accurate due to the difficulty of accessing information, human capabilities or errors. For example, in today's medicine practice, many doctors record patients' medical information (e.g. symptoms, diagnoses, prescriptions, etc) with a audio recorder, and hire external companies to convert recordings into digital records. In the process, an adversary may steal the recording and learn detailed medical conditions of a patient, however, he may learn inaccurate information about patient's identity (e.g. he may not be able to get the correct spelling of the patient's name from doctor's voice). One may assume that the inaccuracy of attackers' knowledge may bring difficulty for them to compromise user identity or privacy. Unfortunately, such inaccuracy could be corrected by collaborating with external information sources, and the privacy risks causes by such attacks should no longer be ignored.

Here is a simple but representative example: Dr. Bob treats Alice in the hospital, while Malory eavesdrops the conversation, or peeps the record. Malory possesses the full prescription with an inaccurate version of Alice's last name (due to Dr. Bob's squiggling handwriting). Mallory does not know Alice, so he starts his attack by first looking into the phonebook for all "similar" names in the neighborhood. The question is: *What is Malory's opportunity of accurately recovering Alice's full name*?

To further articulate this problem, we define *k-approximate-anonymity* as follows:

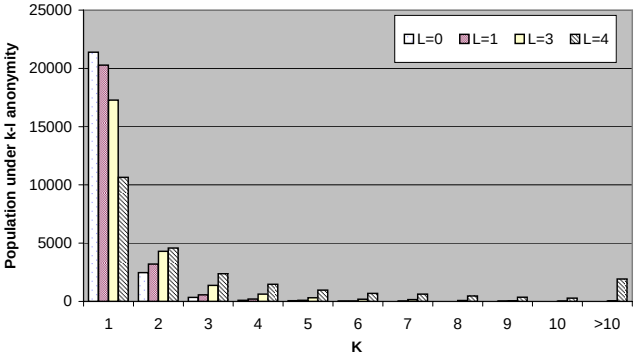

Figure 2: Population under *k-l-anonymity*.

Definition 1 (k-approximate-anonymity) *Given a dataset D, and a distance function $dist(r_1, r_2)$ that returns the distance for any two records on the dataset; for any record r, if there exists $k-1$ records r_x that $dist(r, r_x) <= l$ where l is a preset threshold, we conclude that D satisfies k-approximate-anonymity or k-l-anonymity with dist.*

In the above definition, when $l = 0$, it becomes the original *k-anonymity*. It basically says that when Mallory possesses approximate information on a target, he cannot distinguish the target from $k - 1$ other records in the database.

To simulate the above scenario, we have designed an experiment to study the identifiability of real names in the presence of inaccurate information from the attackers. We first implement a crawler to download the public residential phone book. In a few days, it successfully collects 24399 records from State College area, which covers approximately 64% of the population (according to 2000 cencus data). In each record, we have phone number, first and last names, and full residential address. In the experiments, we use full name as identifiers, and use the Levenshtein distance (edit distance) [17] as the distance function. For different threshold l, we show the population whose names are protected under *k-l-anonymity* in Figure 2.

From the figure, we can see that, with larger l, people are less identifiable with their names. However, overall, most (more than 70%) people are uniquely identifiable even when LD=2, and . It means that even though Malory gets an inaccurate name of the target, he has a good chance to correct the mistake and limit the target to a small range with the help of digital phonebooks. Even when Malory gets four letters wrong in the name, in more than 80% of the cases, his target is limited to no more than 5 candidates, i.e. he only needs to further examine no more than 5 records to identify the target. As we expected, people with longer names or unusual names are more vulnerable, while people with shorter or more popular names are less identifiable, especially when the attacker possesses inaccurate information.

5. CONCLUSION

In this position paper, we study the privacy vulnerabilities when medical records join with the Web. First, we show that multiple information sources (e.g. social networks and public records) could be utilized by the attackers. With attribution, inference and aggregation attacks, the attacks are capable of reconstructing very comprehensive user profiles, with various types of highly sensitive and private information (e.g. names, phone numbers, birth dates, diseases, lab test results, etc). On the other hand, we show that people are very identifiable if the attackers are equipped with information retrieval and data mining techniques. Even though an attacker only possesses a piece of inaccurate information, he is still highly likely to identify the target with the help of external information sources.

6. REFERENCES

[1] L. A. Adamic and E. Adar. How to search a social network. *Social Networks*, 27(3):187–203, 2005.

[2] G. J. Annas. Hippa regulations: a new era of medical-record privacy. *The New England Journal of Medicine*, 348(15):1486–1490, 2003.

[3] S. Barnes. A privacy paradox: Social networking in the united states. *First Monday*, 11(9), 2007.

[4] J. Caverlee and S. Webb. A large-scale study of myspace: Observations and implications for online social networks. In *Proceedings of the International Conference on Weblogs and Social Media*, 2008.

[5] B.-C. Chen, R. Ramakrishnan, and K. LeFevre. Privacy skyline: Privacy with multidimensional adversarial knowledge. In *VLDB*, pages 770–781, 2007.

[6] K. E. Emam. Heuristics for de-identifying health data. *IEEE Security & Privacy*, 6(4):58–61, 2008.

[7] C.-J. Hsiao, P. C. Beatty, E. S. Hing, D. A. Woodwell, E. A. Rechtsteiner, and J. E. Sisk. Electronic medical record/electronic health record use by office-based physicians: United states, 2008 and preliminary 2009. National Ambulatory Medical Care Survey, Dec 2009.

[8] C. Lampe, N. Ellison, and C. Steinfield. A face(book) in the crowd: social searching vs. social browsing. In *CSCW '06*, 2006.

[9] N. Li, T. Li, and S. Venkatasubramanian. t-closeness: Privacy beyond k-anonymity and l-diversity. In *ICDE*, 2007.

[10] B. Luo and D. Lee. On protecting private information in social networks: A proposal. In *ICDE Workshop on Modeling, Managing, and Mining of Evolving Social Networks (M3SN)*, 2009.

[11] A. Machanavajjhala, D. Kifer, J. Gehrke, and M. Venkitasubramaniam. *l*-diversity: Privacy beyond *k*-anonymity. *TKDD*, 1(1), 2007.

[12] D. J. Martin, D. Kifer, A. Machanavajjhala, J. Gehrke, and J. Y. Halpern. Worst-case background knowledge for privacy-preserving data publishing. In *ICDE*, pages 126–135, 2007.

[13] S. Preibusch, B. Hoser, S. Gürses, and B. Berendt. Ubiquitous social networks - opportunities and challenges for privacy-aware user modelling. In *Proceedings of Workshop on Data Mining for User Modeling*, 2007.

[14] C. Safran, M. Bloomrosen, and W. H. et.al. Toward a national framework for the secondary use of health data: an american medical informatics association white paper. *Journal of American Medical Informatics Association*, 2007.

[15] L. Sweeney. k-anonymity: A model for protecting privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 10(5):557–570, 2002.

[16] C. Valli. The insider threat to medical records: Has the network age changed anything? In *SAM 2006*. CSREA, 2006.

[17] R. A. Wagner and M. J. Fischer. The string-to-string correction problem. *J. ACM*, 21(1):168–173, 1974.

[18] P. A. H. Williams. The underestimation of threats to patient data in clinical practice. In *3rd Australian Information Security Management Conference*, pages 117–122, Perth, WA, 2005.