Workshop

INTELLIGENT METHODS FOR PROTECTING PRIVACY AND CONFIDENTIALITY IN DATA

Held in conjunction with the 23rd Canadian Conference on Artificial Intelligence 2010,
May 31st to June 2nd, University of Ottawa, Ontario Canada

PROCEEDINGS

Edited by
Khaled El Emam, Marina Sokolova

Sponsors

NLP TECHNOLOGIES

MultiCorpora

AILIA.ca
ASSOCIATION DE L'INDUSTRIE DE LA LANGUE
LANGUAGE INDUSTRY ASSOCIATION

caiac

PALOMINO
SYSTEM INNOVATIONS INC.

uOttawa

http://www.ehealthinformation.ca/cai/index.asp

Ottawa, Canada
May 30th, 2010

Workshop

INTELLIGENT METHODS FOR PROTECTING PRIVACY AND CONFIDENTIALITY IN DATA

PROCEEDINGS

Ottawa, Canada
May 30th, 2010

ISBN 978-0-9866482-0-5

Workshop Organizers

Khaled El Emam
Children’s Hospital of Eastern Ontario Research Institute and University of Ottawa, Canada
Marina Sokolova
Children’s Hospital of Eastern Ontario Research Institute, Canada

Workshop Program Committee

Dr. David Buckeridge, McGill University, Canada
Nigel Collier, National Institute of Informatics, Japan
Bradley Malin, Vanderbilt University, US
Joel Martin, National Research Council, Canada
Stan Matwin, University of Ottawa, Canada
Dr. Dimitar Tcharaktchiev, The Medical University, Sofia, Bulgaria
Dr. Karen Tu, Institute for Clinical Evaluative Sciences and University of Toronto, Canada

Technical Editors

Elizabeth Jonker, Children’s Hospital of Eastern Ontario Research Institute and University of Ottawa, Canada
Jennifer Noseworthy, Children’s Hospital of Eastern Ontario Research Institute and University of Ottawa, Canada

Table of Contents

Introduction

With the increasing adoption of electronic medical/health records and the rising use of electronic data capture tools in clinical research, large electronic repositories of personal health information (PHI) are being built up. At the same time, large medical data breaches are becoming common. Data breaches may be caused by errors committed by insiders at the data custodian sites, or by malicious insiders. Data breaches can also be caused by outsiders breaking into the data repositories. These data breaches represent legal and financial liabilities for the data custodians, and erode public trust in the ability of data custodians to manage their PHI.

An area that has grown in importance to manage the risks from breaches is data leak prevention (DLP). DLP technologies monitor communications or networks to detect PHI leaks. When a leak is detected the affected individual or organization is notified, at which point they can take remedial action. DLP can prevent a PHI leak or detect it after it happens. For example, if DLP is deployed to monitor email then a PHI alert can be generated before the email is sent. If DLP is used to monitor PHI leaks on the Internet (e.g., on peer-to-peer file sharing networks or on web sites), then the alerts pertain to leaks that have already occurred, at which point the affected individual or data custodian can attempt to contain the damage and stop further leaks.

Computational AI is a key enabling technology for next-generation DLP technologies. This workshop aims to bring together researchers working on computational tools for DLP.

Panel: The Reality of DLP for Health Care Providers

The Panelists

Khaled El Emam is an Associate Professor at the University of Ottawa, Faculty of Medicine and the School of Information Technology and Engineering, a senior investigator at the Children's Hospital of Eastern Ontario Research Institute, and a Canada Research Chair in Electronic Health Information at the University of Ottawa. His main area of research is developing techniques for health data anonymization. Previously Khaled was a Senior Research Officer at the National Research Council of Canada, and prior to that he was head of the Quantitative Methods Group at the Fraunhofer Institute in Kaiserslautern, Germany. He has (co)-founded two companies to commercialize the results of his research work. In 2003 and 2004, he was ranked as the top systems and software engineering scholar worldwide by the Journal of Systems and Software based on his research on measurement and quality evaluation and improvement, and ranked second in 2002 and 2005. He holds a Ph.D. from the Department of Electrical and Electronics, King's College, at the University of London (UK). His lab's web site is: http://www.ehealthinformation.ca/.

Anne Lavigne is the Privacy Officer at The Ottawa Hospital (TOH) and at the Institute for Clinical Evaluative Science (ICES)@uOttawa . She is an experienced hospital administrator, who has worked in the health care field for over 12 years. Anne obtained her CIPP/C certification, the first national data protection certification, and is completing her Information Access and Protection of Privacy certificate from the University of Alberta. In 2004, in order to comply with the Personal Health Information Protection Act, TOH introduced a new Privacy Program. Anne played a pivotal role in the development and implementation of the program at TOH.

Today, Anne is responsible for TOH and ICES@uOttawa's privacy strategies and ensuring compliancy with the Act. Anne also plays an active role in regional privacy initiatives. TOH's Privacy Program was recognized by the Health Care Public Relations Association, Canada's professional association for healthcare communicators when they honored TOH's Privacy Video with the Excellence in Health Care Communications award.

Tyson Roffey is currently the Chief Information and Privacy Officer at CHEO (Children's Hospital of Eastern Ontario). Prior to joining CHEO in October 2007, Tyson was the Senior Director Business Development, Bell Centre for Healthcare Innovation. Among his most recent accomplishments, Tyson has led the strategy, business development and solution architect teams in the creation of a new IS solution for a national service provider supporting health care clients. Tyson's leadership skills and strong track record in innovation, development of IS solutions, and business transformations will prove indispensable to CHEO.

A Systematic Approach to PHI Leak Prevention in Continuous Health Care Data Integration

Jun Hu, Liam Peyton, Khaled El Emam

SITE, University of Ottawa, Canada
{jhu045,lpeyton,kelemam}@uottawa.ca

Abstract. With widespread use of the Internet, knowledge discovery that integrates data across all sources in a health care network on a continuous basis would be useful for disease surveillance and performance management. However, the potential for Personal Health Information leaks is a serious concern, both for ethical reasons and because privacy legislation controls the use of personal health information. This paper presents a systematic approach to preventing leakage of Personal Health Information based on de-identification in order to support continuous health care data integration. Two types of data integration, aggregation and record linking, are considered with two different types of de-identification, anonymization and federated pseudonyms.

Keywords: Personal health information (PHI), PHI leak prevention, data integration, de-identification, anonymization, federated pseudonym.

1 Introduction

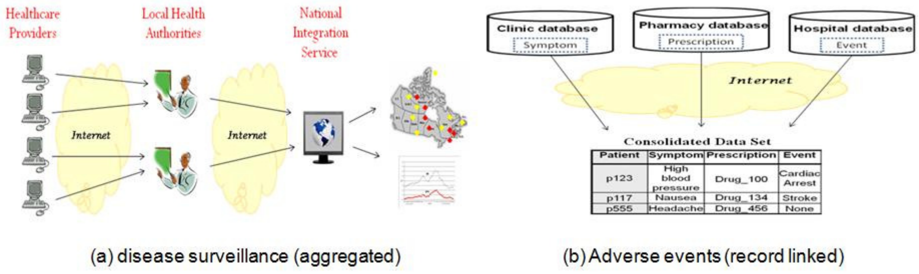
Knowledge discovery that integrates data across all sources in health care network on a continuous basis is useful for disease surveillance and performance management. Privacy legislation [1, 2] controls the use of personal information, and especially personal health information (PHI). PHI leaks are becoming a major security issue when sharing and integrating health care data. An effective approach to reducing PHI leaks is de-identification in which personal identifying information is separated from health information. There are a wide variety of techniques and protocols that can be employed to de-identify data in order to allow knowledge discovery, depending on the requirements of the situation, the level of trust among stakeholders, and the potential risk of re-identification [6]. Federated identity management [10, 16], access control [9], encryption [11], and intelligent monitoring [15] are all relevant to preventing PHI leaks by de-identification in a manner that enables data integration to support knowledge discovery.

However, for most health care data custodians, it is complex and challenging to balance the benefits of knowledge discovery against the risks of PHI leaks, especially if knowledge discovery requires data sharing with other organizations across a network. A systematic approach is needed for PHI leak prevention in the context of continuous health care data integration that addresses methodology, architecture, and identity management as well as detailed protocols. In previous work [7], we proposed both a systematic methodology and an Internet data integration architecture for privacy-protected knowledge discovery in a B2B network by revising and extending the CRISP-DM [4] standardized methodology for knowledge discovery.

In this paper, we further refine the approach by identifying two different types of data integration, aggregation versus record linking, and matching them with two different types of de-identification, anonymization techniques [12, 13, 14], and federated pseudonyms [3]. We define the key criteria or concerns that our approach addresses in preventing PHI leaks while supporting continuous data integration. For each type of data integration, we use a case study to illustrate how to address them based on our approach with the appropriate techniques and protocols.

2 Criteria for Privacy-Enhanced Data Integration

In this section, we will identify a set of criteria or concerns that must be considered if we are going to have a common approach to prevent PHI leaks in supporting continuous data integration. The criteria will be relevant to both types of data integration (aggregation versus record linking). Examples of both types are shown in Fig. 1.   One involves disease surveillance (e.g. H1N1) by aggregating case counts, and the other involves detecting prescription adverse events through linkage of patient records



Consolidated Data Set

Patient	Symptom	Prescription	Event
p123	High blood pressure	Drug_100	Cardiac Arrest
p117	Nausea	Drug_134	Stroke
p555	Headache	Drug_456	None

(a) disease surveillance (aggregated)

(b) Adverse events (record linked)

Fig. 1 Aggregated vs. Record Linked Data Integration

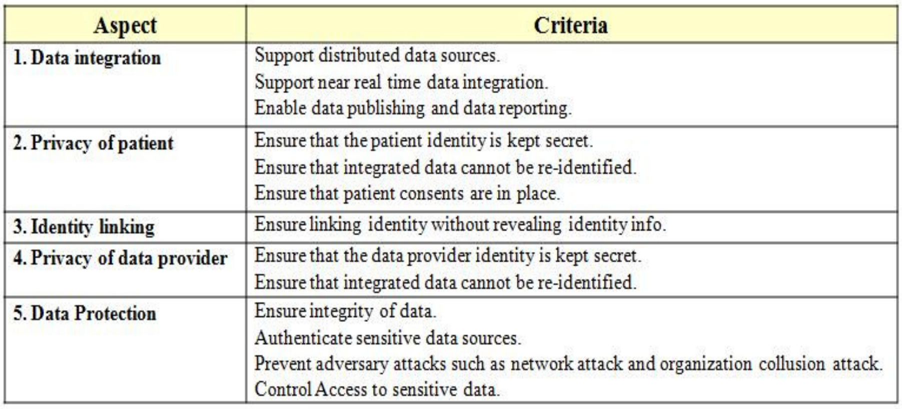
In disease surveillance, data about H1N1 patients is aggregated across health care providers and local health authorities by a national service in order to track the progress and severity of the outbreak across different geographical regions. The issues faced here involve not just the privacy of the patient, but also protecting the privacy and reputation of health care providers who might otherwise be unwilling to share the data. Because we are only interested in aggregated totals there is no need to link patient records across different health care providers.

In adverse event tracking, however, we are trying to find evidence of prescription drug adverse events by correlating and linking patient specific information across pharmacies, hospitals and doctor clinics. The issues faced here are not just privacy of the patient, but also linking patient identity while preserving privacy.

Table 1 shows the aspects and specific criteria that need to be addressed in a systematic approach to eliminating PHI leaks while still supporting continuous integration of health care data for knowledge discovery. The two basic aspects of the approach that must be addressed is to support data integration while ensuring data is protected. For example, the national integration service should be able to calculate and report the H1N1 counts in different strata from the data submitted by healthcare providers across Canada, but neither it nor the local health authorities know how many patients in a particular provider. And the system should ensure that the data is submitted from a real healthcare provider but not a fake provider; and ensure that the data is correct and not modified

More specifically, the approach must provide guidance in how to link data records based on patient identity while still protecting both the privacy of the patient and the privacy of the health care data provider so that data can be shared confidentially within in a health care B2B network. For example, the data integration service should allow researchers to conduct correlation analysis based on a composite view of a single patient whose data is located in multiple data providers; but neither researchers nor data providers know who the patient is

Table 1 Criteria for privacy-enhanced continuous data integration to prevent PHI leaks



Aspect	Criteria
1. Data integration	Support distributed data sources. Support near real time data integration. Enable data publishing and data reporting.
2. Privacy of patient	Ensure that the patient identity is kept secret. Ensure that integrated data cannot be re-identified. Ensure that patient consents are in place.
3. Identity linking	Ensure linking identity without revealing identity info.
4. Privacy of data provider	Ensure that the data provider identity is kept secret. Ensure that integrated data cannot be re-identified.
5. Data Protection	Ensure integrity of data. Authenticate sensitive data sources. Prevent adversary attacks such as network attack and organization collusion attack. Control Access to sensitive data.

3 A Systematic Approach to PHI Leak Prevention

There is no one protocol, algorithm or technique that a network of health care data providers can embrace to ensure PHI leak prevention is managed and balanced appropriately against the needs of knowledge discovery. There are competing interests between the two objectives that must be balanced on a case by case basis with proper consideration given to the objectives of organizations, patients and legislators. However a systematic approach can be taken to designing and selecting appropriate protocols and techniques within the context of a general methodology and architecture. As shown in Fig. 2, our approach is grounded first in a methodology and general architecture [7] that must be agreed to and followed by all members of the health care network. Within that context, we can distinguish two types of approaches to identity management that will guide the selection of appropriate techniques and protocols. The approaches in the right side of the diagram based on anonymization are generally good for aggregation (see Fig.1 (a)), while those in the left hand side based on federated pseudonyms which are generally good for record linkage (see Fig.1(b)). Both approaches have a similar architecture and follow the same methodology, but each takes a different approach to identity management and consequently requires a different data integration protocol.

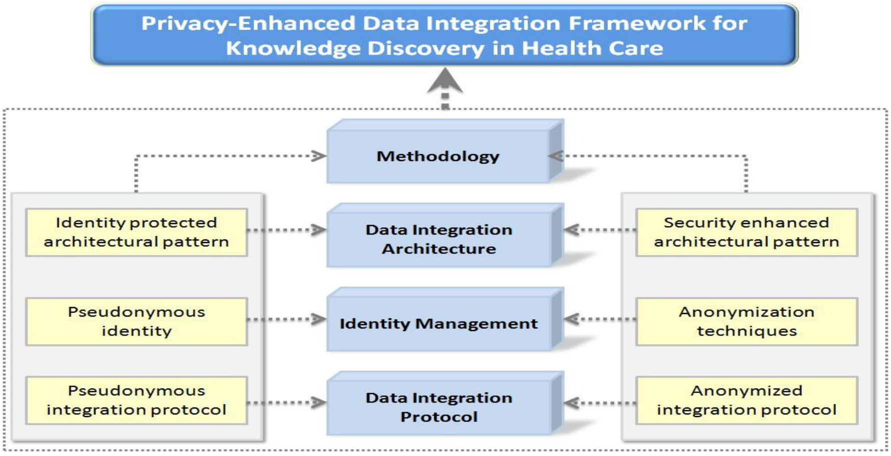

Privacy-Enhanced Data Integration Framework for Knowledge Discovery in Health Care
Methodology
Identity protected architectural pattern
Data Integration Architecture
Security enhanced architectural pattern
Pseudonymous identity
Identity Management
Anonymization techniques
Pseudonymous integration protocol
Data Integration Protocol
Anonymized integration protocol

Fig. 2 A Systematic Approach to PHI Leak Prevention

3.1 PHI Leak Prevention by Anonymization Techniques for Aggregation

In anonymized data integration, there are many anonymization techniques, which can be classified into three categories: data reduction, data perturbation and encryption. The commonly used protocols include data aggregation, k-anonymity [17], anonymized data linking [13], and secure multiparty computation [14]. To facilitate H1N1 surveillance in Fig. 1(a), we designed a k-key holder aggregation protocol to obtain counts as shown in Fig. 3. The k-key holder aggregation protocol adopts the threshold Paillier cryptosystem [11, 12]. Each data provider submits encrypted counts to the aggregator, who sums up the encrypted count by group of at least n providers, and then sends the resulting counts to all k key holders. Each key holder decrypts the aggregated encrypted counts and sends this partial encryption to the data integration service that combines all partial encryptions and obtains the real aggregated count.

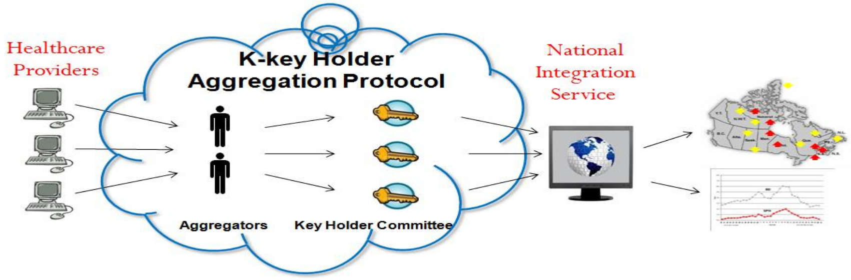

Healthcare Providers
K-key Holder Aggregation Protocol
National Integration Service
Aggregators
Key Holder Committee

Fig. 3 K-key holder aggregation protocol for N1H1 surveillance

A prototype system has been implemented using two aggregators and three key holders. Encryption keys of size 215-bit, and (2, 3)-threshold version of Paillier algorithms [11, 12, 14] were used. A simulation of 3000 data providers and 200 regions showed that the calculation time in aggregators, key holders and data integration service and the number of data providers and regions were linear. The performance was reasonable and the protocol was scalable.

This approach addresses all the criteria from Table 1 (except identity-based linkage which is not required). In particular, aggregators do not know totals for each provider (important for criteria 4) and k-key holders reduce the risk of collusion (criteria 5).

3.4 PHI Leak Prevention by Pseudonyms for Record Linking

With federated pseudonyms, one can link events across sessions to a pseudonym identity created for each organization, without knowing the actual identity of the patient. To implement this for a specific health care network, one would create a Master Patient Index maintained by an identity provider. The identity provider manages and protects the identity of patients separate from their health care data. Each health care data provider has their own pseudonym for a patient in order to manage their health care records, but data providers are not able to link records between themselves, since each has a different pseudonym for the patient given to them by the identity provider.

In Fig. 4, data can be securely linked between data providers through the mediation of the identity provider and an additional separate data integration service. The data integration service, through its interaction with the identity provider requests the data sets it desires from each data provider. The datasets are returned in two pieces. The pseudonym attributes for each row are returned by the identity provider which transforms them into pseudonyms meaningful to the data integration service using the Master Patient Index (MPI). The remaining attributes are sent to the data integration service by each data provider.

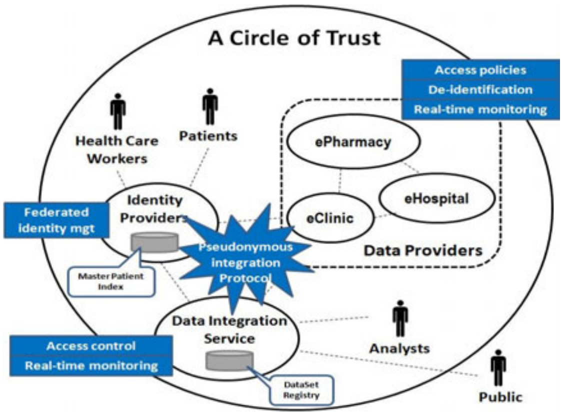


Fig. 4 Pseudonymous Data Integration in a Liberty Alliance Circle of Trust

This case study has been described in detail in a previous paper [8]. This approach addresses all criteria from Table 1, especially the need for identity linkage. A key component is the Dataset Registry that registers datasets used and created by the integration service, including related business agreements and access controls, as well as their status as they are created according to our methodology [7]. It maintains an audit trail [5] monitoring all access requests. Both integration services in Fig. 3 and Fig. 4 can be supported by a Dataset Registry.

Another key aspect of both approaches is the introduction of third party components and organizations in a network architecture that take responsibility for

a) identity management and linking separate from data providers so that health care data is de-identified
b) data integration and publishing separate from the data providers who provide the source data

4 Conclusion

PHI leaks pose a significant and increasing problem for organizations, especially in B2B healthcare networks where one wants to balance the need for optimal privacy protection with the opportunity for knowledge discovery. A systematic approach to preventing PHI leaks should be based on defining appropriate architectural patterns or frameworks that separate and manage identity information from health information. Depending on the requirements and chosen identity management, different protocols and techniques can be bundled with the architecture to provide different types and levels of protection.

Our approach provides a mechanism for organizations to systematically address the major criteria and concerns based on a common methodology and architecture. We have shown how two very different types of data integration can be done securely within a health care B2B network based on a this common methodology and architecture where organizationally identity management, linking and protection is handled by a third party separate from the source data providers and where data integration and publishing is also handled by a different third party separate from the source data providers.

References

1. HIPAA, Health Insurance Portability and Accountability Act, United States Congress, United States, 1996. http://aspe.hhs.gov/admnsimp/pl104191.htm Accessed April 2010.
2. PIPEDA, The Personal Information Protection and Electronic Documents Act, Department of Justice, Canada, 2000. http://laws.justice.gc.ca/en/P-8.6/text.html Accessed April 2010.
3. The Liberty Alliance, http:/www.projectliberty.org/ Accessed April 2010.
4. Shearer, C.: The CRISP-DM Model: The New Blueprint for Data Mining, Journal of Data Warehousing, 5(4), pp. 13-22 (2000)
5. Peyton, L., Hu, J., Doshi, C., Seguin, P.: Addressing Privacy in a Federated Identity Management Network for E-Health, 8th World Congress on the Management of eBusiness, Toronto (2007)

6. El Emam, K., Jabbouri, S., Sams, S., Drouet, Y. & Power, M: Evaluating Common De-Identification Heuristics for Personal Health Information, Journal of Medical Internet Research, 8(4):e28 (2006)
7. Hu, J., Peyton, L.: A Framework for Privacy Assurance and Ubiquitous Knowledge Discovery in Health 2.0 Data Mashups, Ubiquitous Health and Medical Informatics: The Ubiquity 2.0 Trend and Beyond (Eds. Sabah Mohammed and Jinan Fiaidhi). ISBN13: 9781615207770. IGI Global (2010)
8. Hu J., Peyton L., Turner, C., Bishay, H.: A model of trusted data collection for knowledge discovery in B2B networks. In Proceedings of the 2008 International MCETECH Conference on e-Technologies, pp. 60-69. Montreal, Canada. (2008).
9. Sandhu, R. S., Coyne, E. J., Feinstein, H. J., and Youman, C. E.: Role-based access control models". IEEE Computer Vol.29, No.2, Feb., p38–47 (1996)
10. Koch, M., & Möslein, K.M.: Identity Management for Ecommerce and Collaborative Applications. International Journal of Electronic Commerce, 9(3), 11–29 (2005).
11. Paillier P.: Public-key cryptosystems based on composite degree residuosity classes. EUROCRYPT'99. (1999)
12. Fouque P-A., Poupard G., and Stern J.: Sharing decryption in the context of voting or lotteries. In Proceedings of Financial Cryptography '00, LNCS 1962, Springer-Verlag (2000)
13. Swire,P.: Research Report: Application of IBM Anonymous Resolution to the Health Care Sector. http://www.ehcca.com/presentations/cclf3/swire_s5_t4.pdf Accessed April 2010
14. Kantarcioglu, M. et al.: Formal anonymity models for efficient privacy-preserving joins, Data Knowl. Eng. (2009)
15. Sokolova, M., El Emam, K., & et al: Personal health information leak prevention in heterogeneous texts. Biomedical Information Extraction International Workshop, held jointly with the 7th International Conference on Recent Advances in Natural Language Processing. (2009).
16. Peyton L., Hu, J.: Federated Identity Management to Link and Protect Healthcare Data. International Journal of Electronic Business (IJEB). 8(3). (2010).
17. El Emam, K., Dankar, F.: Protecting privacy using k-anonymity. Journal of the American Medical Informatics Association, September/October, 15:627-637 (2008)

A Brief History of Inadvertent Sharing on P2P Networks: Causes, Current Solutions and Future Directions

Nathan Good
828 San Pablo Ave Suite 120D, Albany CA 94706
nathan@goodresearch.com

Abstract. This paper discusses a brief history of inadvertently shared sensitive personal information on P2P networks and the measures taken to correct these issues and protect consumers. Despite these measures, rapid changes in technology and the challenges of designing for complex consumer behaviors are discussed as possibilities for the continued existence of private information on these networks. Inadvertent sharing on P2P networks reveals larger issues regarding the design of secure and private networked systems for consumers sharing and storing their personal information on networked devices. This paper proposes addressing the larger issue of usable security and privacy through focusing on usability concerns and a deep understanding of consumer's behavior to better target technological solutions and policy goals.

Keywords: P2P, inadvertent sharing, usability, security, privacy, personal information

1 Introduction

Inadvertent sharing of personal files over P2P networks continues to be of interest to policy and technical communities, both because of the types of personal information available on the network[1,4,5,6,7,8,10,12] and how accessible it is for any one to access this sensitive information with simple keyword searches and without any specialized expertise. Inadvertent sharing has received attention from regulatory agencies as well[6], and is the subject of a bill currently proposed in the United States [13].

As discussed in previous work[4,7,9], the reason for private files being exposed inadvertently is largely due to a combination of human factors and issues arising from user interface design. For example, a user who is looking for video files may assume incorrectly that only video files on his network can be shared. Also, a user may not understand that when a folder is added to be shared with a P2P network, all folders contained in that folder are shared as well.

Early work [4] uncovered design issues that contributed to inadvertent sharing, such as using wizards to automatically add files, recursively adding folders and by default sharing all file types. Some of these suggestions have been incorporated into industry best practices[3] and adopted by some of the larger, more popular P2P clients. Additionally, aggressive technical measures have been taken by ISPs, AV firms, and others to limit the number of clients (quarantine, site-blocking, filtering, etc) and bandwidth they have available, as well as track information shared over their networks. In addition to these measures, many companies prevent users with sensitive information on their computers from installing these applications, or have policies in place that prohibit their installation.

Despite these measures, reports still come in about sensitive personal information accessible on these networks[1,4,5,6,8,10,12]. Also, despite this widespread interest and knowledge of the types of sensitive information shared on these networks, this paper argues that more could be done to understand the specifics of why this sensitive information still persists. Understanding these factors is important to realistically address current P2P issues. Lessons learned from P2P can be applied to design the many kinds of systems in the near future that consumers will use to store and share sensitive information with others (for example health records).

1.1 Brief history of P2P and inadvertent sharing

One of the earliest consumer P2P application was Scour Media Agent. Scour was a program that "scoured" the internet for computers with open window shares (SMB shares) and indexed files on those shares to share with others. Scour, unlike Napster, indexed image and video files as well as music and indexed all shares it could find on the web, and not just the shares of its users. Scour was perhaps the first example of inadvertent file sharing on P2P. However, in this case it was the UI and defaults for Microsoft Windows SMB sharing that caused confusion among users, which Scour used to populate its index.

Future P2P products, such as Napster, continued to focus mainly on music or mp3 files, sometimes including wav or midi formats as well. It was not until KaZaA that any type of file was available over P2P networks. KaZaA developed a new more robust architecture including "super peers", and also created a popular client that allowed sharing of any file type over the network.

As P2P moved away from a specific file type to a more general file types, the risk of sharing of personal information grew. Not surprisingly, it was the first client to allow sharing of any file type that was perhaps the first to have inadvertent sharing of personal information. Good et al[7] reported on how usability flaws in the KaZaA application could lead to end users sharing their personal information, including their whole hard drive, without being aware of what they were sharing. When the concept of any file type was added, the users "mental models" broke down and without the aid of the UI or a thorough understanding of their file system there was a possibility that they would inadvertently share personal files.

Subsequent protocols including Gnutella and Bit Torrent, resulted in many different popular clients, including Limewire, uTorrent, with different interfaces. BitTorrent was unlike previous P2P file sharing programs in that it required the creation of a metafile for the file to be shared, and then this meta file needed to be posted for others to see. Because of these steps, it was less likely to be prone to inadvertently sharing personal information than other protocols.

Gnutella remained a traditional P2P protocol, and the most popular Gnutella client was Limewire. Limewire allowed users to share any kind of file over its network.

Consequently, similar issues with private information being shared arose on LimeWire as well. As the popularity of previous clients such as KaZaA faded, Limewire became the largest P2P client, and the most targeted for issues of inadvertent file sharing.

2 Current Issues

2.1 Addressing Inadvertent Sharing

Since the initial discussion of inadvertent file sharing, there has been a large amount of activity on the policy and technology side to address concerns with private information on P2P networks. In the United States, there has been three congressional hearings on inadvertent file sharing and a workshop by the Federal Trade Commision, in addition to pending legislation [13]. The US FTC has reported on business use of P2P applications and risks of inadvertent sharing and many popular anti-spyware and anti-virus programs will block and quarantine P2P programs as well.

Industry groups such as the DCIA have been the creating best practices[3] that require their members, including programs such as LimeWire, to take steps in the way they design their user interfaces to prevent inadvertent sharing in order to be compliant.

On the user interface side, some of the more popular companies have changed their user interface to not allow sharing of popular documents by default, and included warnings and require explicit actions in order for users to share specific files. The latest version of LimeWire for example has significantly revamped their user interface. It prevents default sharing of sensitive file types, separates the download from the shared folders, and now requires confirmation and affirmative actions on the part of the user in order to recursively add new files or folders to be shared. Files described as sensitive file types must be shared explicitly after checking several boxes, and users can readily see what is being shared from one screen.

Technology changes rapidly, however, and with that the landscape of popular P2P clients has evolved as well. Formally dominant players as such as KaZaA and Morpheous have given way to Limewire, Frostwire and others.

Finally there has been a shift in file sharing protocols to BitTorrent which require explicit steps to create torrent files and reduce the likelihood of sharing files inadvertently. In addition, files are repeatedly being hosted on one click hosters such as megaupload, where users explicitly upload content to be shared.

2.2 Why is inadvertent sharing still a problem?

Despite the attention it has received, as well as the security mechanisms put in place to prevent and track inadvertent files, Recent literature suggests that a variety of private information exists on P2P networks [4,8,9]. While precise reasons for inadvertent file sharing are questions for further research, there are several factors that could contribute to the continued prevalence of private documents persisting on P2P networks. These are discussed below in terms of user behavior, technology trends.

2.2.1 Factors with user behavior

Changes in popular clients and UIs

One estimate is that there are over 250 existing P2P programs today[5]. As the types of clients change, so do the user interfaces designed for them. The user interfaces for new or upcoming clients may or may not choose to use best practices that will likely reduce a users ability to inadvertently share files. For example, a new program needs to decide if it will require that sensitive file types are not shared by default.

Multiple users per machine.

Many anecdotes around inadvertent sharing of personal information mention cases where a computer is used by more than one user. Users share accounts and share logins, and consequently may end up sharing personal information. In many cases, the software assumes that the user who has configured the application is the one using it. One challenge is to design the software so that even a user who is not familiar with it can readily identify it is running, what is being shared and how to turn it off.

Blurring of the line between home and work

People share computers and files between work and home in order to get their work done. This makes it easy for them to accomplish their goals, and encryption, access controls and other technologies are still not easy or common enough for end users to use efficiently and seamlessly in their daily work.

Increase in personal digital libraries and sensitive digital information

People are keeping more personal and sensitive information in digital form, are unable to keep track of it, and consequently organize or delete it.

Convenience vs Security

Free tools available online are often of higher quality and much easier to use than proprietary tools designed to protect security. For example, many users find it easier to send a copy of document through gmail in order to view it on another computer rather than attempt to transfer it through the network. Consequently, many copies of documents exist in different places, and are vulnerable.

3 Future Work & Conclusions

It may be the case that users are much more willing to share now than they were previously, and this opens them up additional threats. As the popularity of websites such as 23andMe.com and PatientslikeMe. com demonstrated, users want to share personal, even sensitive information with pseudo-strangers. Additionally, they want to share work with others, they want to share photos, they want to share other items. As their personal lives become more digital, they will naturally want to store and share these intimate materials with others, and this opens them up to potential risks. Current issues with P2P and inadvertent sharing highlight the challenges of providing the right policies, technologies and designs to allow sharing of information while providing users optimal control over their experience. In all cases there are potential areas for abuse. For example, a user interface that actively deceives users and shares sensitive information should not be allowed, however policies that micromanage users interactions also lead to workarounds that can be insecure.

Future studies that combine percentages of exposed files with client types and changes in the user interface can help determine what areas need the most attention.

4 References

1. Claburn T. File sharing exposes supreme court justice's personal information. InformationWeek. 2008 10 July.
2. Compliance Report http://www.dcia.info/activities/ispg/ISPG_Compliance_Report.pdf (2009)
3. DCIA Best Practices http://www.dcia.info/activities/ispg/inadvertentsharingprotection.pdf (2008)
4. Emam, K. E., Neri, E., Jonker, E., Sokolova, M., Peyton, L., Neisa, A., Scassa, T: The Inadvertent Disclosure of Personal Health Information Through Peer-to-Peer File Sharing Programs. Journal of the American Medical Informatics Association. 17, 148—158 (2010)
5. File Sharing Software Reveal Users' Private Information http://www.livescience.com/technology/file-sharing-p2p-private-information-100316.html (2010)
6. FTC findings for new information http://www.ftc.gov/opa/2010/02/p2palert.shtm
7. Good N, Krekelberg A. Usability and privacy: A study of Kazaa p2p file-sharing. Proceedings of the SIGCHI Conference on Human Factors in Computing Systems; 2003.
8. Johnson E, Dynes S. Inadvertent disclosure - Information leaks in the extended enterprise. Proceedings of the Sixth Workshop on the Economics of Information Security; 2007.
9. Johnson E, McGuire D, Willey N. Why file sharing networks are dangerous ? Communications of the ACM. 2009;52(2):134-8.
10. Sensitive Data Leaking Onto P2P Networks http://www.computerworld.com/s/article/299890/Sensitive_Data_Leaking_Onto_P2P_Networks (2007)
11. Smetters, D. K. and Good, N. How users use access control. In Proceedings of the 5th Symposium on Usable Privacy and Security. 2009
12. Vijayan J. Personal data on 17,000 Pfizer employees exposed; P2P app blamed: An employee had installed file-sharing software on a company laptop. ComputerWorld; 2007.
13. Mack, M. B, HR.1319 Informed P2P User Act (2009)

Detection of Personal Health Information in Unstructured Web Documents

Amir Razavi, SITE, University of Ottawa, and CHEO Research Institute
Marina Sokolova, CHEO Research Institute
{arazavi,msokolova}@ehealthinformation.ca

Abstract. We look at studies of Personal Health Information (PHI) leaked through unstructured texts on the Web. We survey the published work and systems used to detect those leakages. We show that some studies concentrate on analysis of personally identifiable information, whereas other studies focus on health information. As a result of such separation, PHI disclosure on the Internet is not studied to a full extent. For future work, we propose a method which can detect PHI leakages in unstructured text. PHI, usually shared on conditions of confidentiality, protection and trust, should not be disclosed to unrelated third parties or the general public.

1 Medical Postings on the Internet

Medical and health information is extensively present on the Internet, in both the traditional media, e.g., online magazines, and the user-generated Web content, e.g., blogs. On line medical journals, wikis, blogs and podcasts are considered being effective educational tools, information recourses for health professionals and a means for raising the awareness of the general public in many different health/medical domains.1 [1,2,3,4,5,6]. Medical blogs, whose authors post entries over time, had become a new connection between health professionals and the public, although there is no precise statistics about the number of these blogs in connection with health care and medical domain [7]. At the same time, some blogs become potentially "hazardous" because of privacy breaches.[8] There are health care professionals who share private beliefs in public areas and take the risk of revealing confidential patient information such as reporting personal experiences, clinical interactions, etc. [9,10,11,12,13,14,15]. The cited studies show that, although medical blogs seem to be a proper forum to share experience and knowledge, they can accurately identify some private aspects trusted by patients to physicians and nurses.

Significant volumes of personally identifiable information (PII) [16,17] and personal health information (PHI) [18] have been found on the Web. In some cases, social workers, doctors, government agencies leaked health information which was disclosed solemnly to them: PHI has leaked from a Canadian provincial government agency through a publicly available website [19] and from health care providers, through documents sent by employees and medical students [20], including notes on treatments and medications taken through the Internet [21].

In this study, we survey work which focuses on the problems posed by PHI leakages on the Web. When information is available, we discuss the applied systems and their core elements. IN Future Work, we propose an approach which can be used to detect PHI in free-form, unstructured text.

2 Personal Health Information

Personal health information (PHI) refers to diseases, symptoms, treatments and other health-specific details of an individual combined with the information which can identify the individual, e.g. person name or date of birth.[22] We divide the personal health information into **personally identifiable information,** which includes Person names, Locations, Addresses, URLs, age-defining Dates (e.g. *Serge, London, Osaka Jasmine, 401 Smyth Rd., Empire State Build., (was born) 02 May 2008, 05/14/07),* and **health information,** which includes Disease names, Symptoms, Drug names, Health care providers, tests and results (e.g. *Pneumonia, Tenosynovitis/ calcium deficiency, labile blood pressure/ Aspirin, Fosamax/ CHEO, Dr. Joe Doe/ RBC, WBC*).

PHI leaks may harm unsuspecting individuals, even if those leaks are inadvertent. For instance, big pharmaceutical companies can be pressuring patients to take expensive drugs that aren't needed, medical malpractice insurance prices have been steeply increased [23,24,25,26,27,28,29,30]. Internet privacy breaches magnify the potential harm because

[1] A wiki is a website that allows the easy creation and editing of any number of interlinked web pages via a web browser using a simplified markup language or a WYSIWYG text editor. A podcast is a series of digital media files (either audio or video) that are released episodically and often downloaded through web syndication. Blogs are articles, diaries, ecce, etc. posted on the Web; their editing, style and language closely resemble those of traditional journal publications.

of the numbers of people involved. A simple act of using a search engine to look for information on a medical condition is much less private than one might imagine. Each piece of PHI potentially can be used/ abused at least by insurance or recruiting companies. [31]

So far, there is a lack of analysis of texts which host both PII and health information, although such texts are primary candidates for PHI leakages . Another understudied question is whether PHI is self-disclosed or published by a presumably trusted confidant (either an individual or a company). For example, a dedicated web site "*YourPrivacy*" separately considers PII in web posts, whereas health is viewed within doctor-patient communication. In the case of PHI disclosure by a trusted party, i.e. a confidentiality breach, a person might not be aware about her PHI leaking before the disclosed information becomes harmful.

3 Studies of Personally Identifiable Information on the Web

Previously, studies analyzed information disseminated by health care professionals-doctor blogs [32] and self-disclosed personal information in relations to stigmatized health conditions (in medical information search) [33,34]. However, those studies did not analyze large volumes of texts. Thus, the published results may not have sufficient generalization power or deliver the prevalence estimates of the leakage [35,36]. For this survey, we performed an exhaustive web search on the topic 'personal information leakages over/through the Web', which included 'personal health information' as a sub-category. We have found several publications related to personally identifiable information in unstructured (Web) text, but no statistics about the PHI leakages were found. The absence of published results is considered as a highly indicative sign that the PHI leakage on the Web is an understudied issue.

We differentiate the found articles into work reporting on personally identifiable information in *general*, i.e. other than medical, domains versus research specifically related to the *medical* domain. The next part of this paper surveys work in *general* domains.

A survey was conducted at the Joint-Research Center of the European Commission for automating event extraction from news articles [37]. The articles were collected through the Internet with the Europe Media Monitor system. Then, Natural Language Processing (NLP)[2] and clustering techniques have been applied to the 30,000 daily news reports. The goal was to identify clusters of similar news items in order to extract the major news items each day. The system first identified the major news articles in each of 13 languages and then located events geopolitically, including the record entities involved. Examples include articles per day for South East Asian countries following the 2004 Tsunami, articles related to the Iranian President: Mahmoud Ahmadinejad in more than 17 different spellings and languages; London Bombers Network which could be investigated through their news reports or through their other relationships; and Automatic VIP identification associated with Pakistan and distinguishing the recent earthquake as the top news. The obtained results showed that those extractions could be used for national comparisons and to derive name variants. In [38], the authors' motivation for automated event tracking was to provide quantitative objective incident data which has a broad coverage of terrorist incidents and violent conflicts around the world, i.e. terrorist attack indicators for Iraq, India, Israel and Indonesia in August 2003. They used those quantitative data to form the basis for further populating the incident databases. The databases were used by systems performing analysis and risk assessment of the terrorist incidents and violent conflicts trend. The system task can be considered as "event extraction". Event extraction systems normally extract data from text to answer some questions like: Who did what to whom, when, where and with what consequences? Any related answer could be regarded as an event and the system should identify as many "facts" as possible from the collection of news articles describing each event.
A body of work compares information revealed by members of social networking sites Facebook[3] and MySpace[4], along with self-disclosed personal information [39]. Although members of both sites reported some similar privacy concerns in the willingness to share their personal information through the development of new relationships, the study suggests that Facebook members expressed greater trust in both the web site and its members, and were more willing to share their identifiable information. We note that these results were manually obtained, thus, may reflect a human bias. In [40], the authors consider relations between the privacy issues and possible victimization of some youth in web-based networking environments (e.g., MySpace). The main expressed concern is that the "possibility of sexual predators and pedophiles finding and then assaulting adolescents who carelessly or unwittingly reveal identifiable information on their personal profile pages" (ibid). They reported that the final publicly accessible youth samples consisted of 1475 profiles that were manually analyzed by their research assistants. Quantitative results, obtained on randomly sampled MySpace profile

[2] NLP is an advanced form of Artificial Intelligence science and technology. NLP deals with unstructured, free-form texts written in language used by humans.

[3] http://www.facebook.com/

[4] http://www.myspace.com/

pages, show types of personally identifiable information publicly posted by youth. The results are revealing: 8.8% of the users posted their full name, 57% included a picture, 27.8% listed their school, and 0.3% provided their telephone number. Based on the gathered evidence, the authors discussed some implications for Internet safety among adolescents. According to them, negative consequences include: cyberbullying, cyberstalking, alcohol and drug abuse, hate crimes, planned or executed bombings, planned school shootings, suicide, and even murder. In addition, the biggest public concern is shown to be a potential vulnerability of youth to online predators and pedophiles.

Privacy risks brought up by the personal information dissemination may require special attention. Such dissemination through project outsourcing was studied in [41]. In the survey, the authors show how specialist IT service providers may play essential roles in transferring specific skills and confidential information to the wider industry context (via leakage). In another study, a web-based Data Leakage Prevention (DLP)[5] system has been proposed which prevents leakages of distributed confidential information [42].

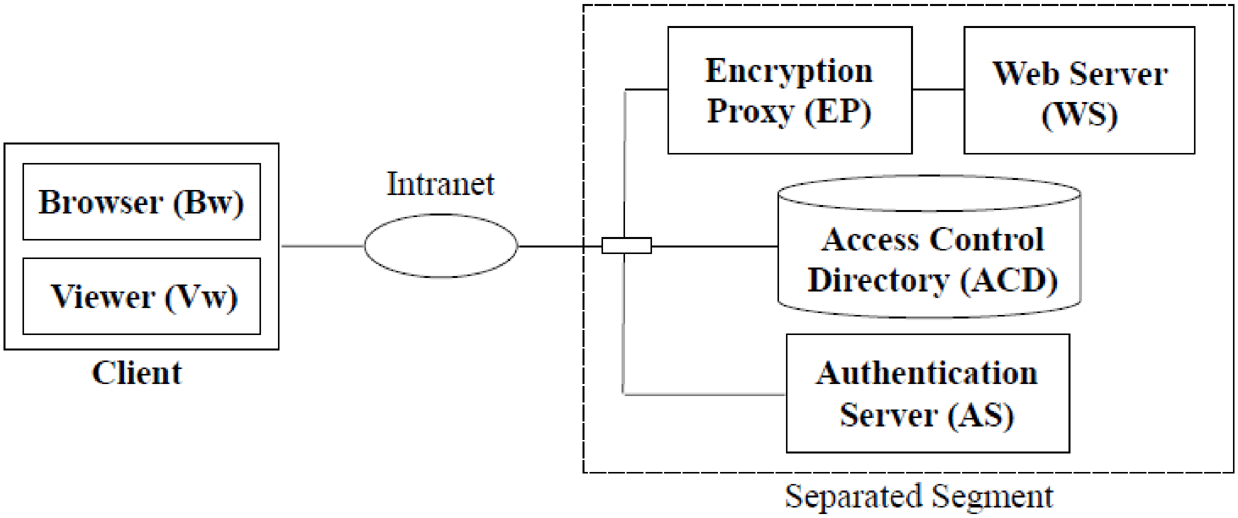


Figure 1. System architecture of the DLP system.

Such leakage can be caused by people who gain the access to identical information shared for different purposes. The system applies the centralized access control (e.g. legal control, bureaucratic control and social control including reputation, concerns, professional ethics and trust) to the distributed confidential information and supports the confidential web pages with the corresponding content key that is a secret key (for encryption and stored in Access Control Directory), dynamically generated by web applications (e.g. customer care web application super-distribution systems).

Many commercial systems are being offered on the software market to prevent information leakages. Among those, NLP/ML -based information leakage prevention systems work on higher levels of the networks. They are noise-resistant, manage large amount of previously unseen information, and are able to process documents written in natural language. Those characteristics designate such systems as strong candidates for prevention of the leakages of confidential information over the Internet. We briefly introduce the systems through *Cut Once*, an extension to Mozilla Thunderbird[6], an open source which implements methods for Email Leak Detection and Recipient Recommendation [ECIR-2008,SDM-2007]. CutOnce, an NLP/ML system, is built to predict and detect sensitive information leakages into the Web. In other words, the system issues some alert in order to prevent the leakages of detected sensitive information into to the Web from the source. CutOnce has to be trained before it is able to make recipient predictions. For instance after the training process it could issue hazardous messages based on the to list, an email or a message is being sent. [43]. Unfortunately, no exact evaluation of the tool's performance is publicly available.

[5] The figure has been extracted from the source work.

[6] http://www.mozillamessaging.com/en-US/thunderbird/

4 Studies of Medical and Health Information on the Web

We observed that human judgement and manual analysis are mainly used for detection of medical and health information breaches and leakages over the Web. In [44], the authors analyzed medical blogs which contain some identifiable information and are likely to be written by physicians or nurses. The study motivation was to examine the scope and content of medical blogs and estimate how often blog authors commented about patients, violated patient privacy, or displayed a lack of professionalism. Researchers applied the Google search term "medical blog" to obtain sampling identifying sites which posted entries from 1/1/ 2006 to 12/14/2006. Five entries per blog were manually reviewed, in order to categorize the characteristics of their contents. The results are stunning. Among 271 identified medical blogs, over half (56.8 %) of blog authors provided sufficient information in text or image to reveal their identities. Individual patients were described in 114 (42. 1%) blogs. Patients were described positively in 43 blogs (15.9%) and negatively in 48 blogs (17.7 %). Among the blogs that described interactions with individual patients, 45 (16.6%) included sufficient information for patients to identify their doctors or themselves. Three blogs showed recognizable photographic images of patients. The Healthcare products were identified, either by images or descriptions, in 31 (11.4%) blogs. The products were included prescription drugs, medical devices and nutritional supplements. The author states that a recent poll showed that 29 percent of bloggers have been approached by a public relations professional to endorse a product, in which 52 percent accepted and had posted endorsements of the products on their blogs. They describe an example in which the anonymous blogger "Flea" revealed details of a patient's death after a malpractice case was carried out against him.

In [45], the authors say that some doctors extensively write their blogs despite complaints about a lack of time to see patients or keep up with the latest innovations and researches. The authors believe that such extensive information may cause some problems for the bloggers are unwittingly revealing confidential patient information. "The blogging community has made an effort to set standards for medical bloggers, but unfortunately, professional organizations and medical educators haven't come out with rules for handling the new medium," (ibid) Tara Lagu of the University of Pennsylvania tells: "Medical blogs are a great opportunity to learn about the health care system, but they need to know some bloggers have unprofessional conduct, although that doesn't represent the medical profession as a whole. The issue is the risk of losing patient trust. We want to maintain that."

5 Conclusions

We found that many studies analyze variety of sensitive Web information related to health. However, the surveyed work shows a lack of verifiable information about PHI detection methods. Many statistical results were obtained based on the manual analysis of the retrieved Web data. Without assessment of the retrieval accuracy, we cannot reliably generalize prevalence of PHI on the Web.

High risk of privacy breaches causes an especial attention to medical blogs which have become an integral part of the public face of healthcare and a new connection between health professionals and the public today. Thus, we support proposal of professional organizations and medical educators to come up with rules for handling the new medium or risk losing the patient trust.

We have shown that some health care professionals inappropriately share privately obtained information in public settings or reveal confidential patient health information. Some PHI leaks are self-disclosed, whereas others result from the disclosure of confidential information leakage caused by people who gain the access to identical information shared for different purposes such as reporting personal experience, clinical interactions, etc. (*e.g.* child's disease symptoms and parent's email and name, wife's medical procedure and husband's employment and name). Although many information leakages are accidental disclosures, there are purposeful betrayals (e.g. by consultants who work for several clients at the same time).

We regard centralized access controls (e.g. legal control, bureaucratic control and social control including reputation, concerns, professional ethics and trust) to be a good approach in prevention of the distribution of confidential information. The listed controls can support the confidential web pages with the corresponding content keys.

We suggest that, for better understanding of PHI leakage problem, Text Data Mining and Information Retrieval tools are well suited than other text analysis methods. NLP/ML -based information leakage prevention systems are rather noise-resistant, manage large amount of previously unseen information, and are able to process documents written in natural language. Those characteristics designate such systems as strong candidates for prevention of the leakages of confidential information over the Internet. Among those systems there are some applicable methods for national comparisons and to derive name variants which could play an essential role for detecting any personal information leaks. Furthermore, event extraction systems can potentially collect some data from text in order to answer some questions like: "Who did what to whom, when, where and with what consequences?" which can be considered as a powerful tool for PHI/PII detection systems.

In our view, human judgement and manual analysis, if correctly implemented, can create a synergetic collaboration with the automated methods. Both the research methodologies are being used for risk assessment and detection of medical and health information breaches and leakages over the Web.

6 Future work

The presented survey has shown that the volume of the PHI leaks on the Web has not been yet empirically determined. Hence, the first and necessary step is to assess the extent of the issue by the means of Text Data Mining and Information Retrieval methods which has been augmented and verified by human judgment. For PHI leakage detection on the Web, we propose a method which uses the advanced learning and text analysis techniques. With the application of those tools, the method finds and organizes the PHI documents into semantically related groups. Then, the extracted PHI semantics will be combined with other document information and used to assess the PHI leakage on the Web.

We propose to apply context clustering (i.e. Multi-parameter Hierarchical Clustering [[46]]), entity recognitions (e.g. Stanford NER and UIMA) and relationship extraction [[47],[48]] as well as some Machine Learning and Natural Language Processing algorithms like SVM classifier or Conditional Random Fields stream modeller. Those techniques can potentially be used in design and deployment of PII and PHI detection systems or in the leak prevention purpose.

Next, we aim to extract document attributes and patterns which are helpful for a faster and more precise exploration of possible PHI leakages. At this stage we are going to search for sites with a higher probability of PHI leakages or breaches. For example, strong candidates for such sites are those which post “Frequently Asked Questions about Living organs Donations”. We can mark those sites as high-risk cases. Then we are going to manually extract some keywords and their patterns through the high-risk cases. Those words and patterns will be used on a data retrieval step in the queries in order to boost the recall of PII/PHI detection on the web. We hypothesize that such an approach will help to expand and diversify the list of identifiers of PHI leaks. As a result, we can use the newly obtained identifiers for better detection of the more diverse PHI.

To generalize those identifiers according to the concepts they represent, we want to apply a novel method which builds multi-layer conceptual text representations. Traditional methods represent all the entries based on their consisting texts, thus, representations are often isolated from the goal. We propose that corpus entries will be onto logically represented based on their context and the goal of the task (PHI detection). The method is called: Text Ontology Representations via Fundamental to Specific Essence (TOR-FUSE).

The multi-layer model will allow data interpretation in a more conceptual space rather than just containing separate words appearing in the text. On all the levels, the model considers the cognitive plausibility of interpretation as one of its basis. It aims to embody knowledge extraction by applying an ontological hierarchy of representation. The hierarchy helps in analysis of a rather large corpus of text based on word order, different configurations of word co-occurrence, etc. [49,50]. The method finds the closeness between words and text passages and applies it to represent text in a more informative way. Hence, TOR-FUSE will be able to extract some latent conceptual bases specifically applicable for the PHI detection task. On the next step, the data representation will be used for training a tailored ensemble of learners in order to extract new patterns for PHI detection.

Acknowledgements. This study has been financially supported by a CIHR operating grant and NSERC research grants.

References

[1] *Boulos MN, Maramba I, Wheeler S. Wikis, blogs and podcasts: a new generation of Web-based tools for virtual collaborative clinical practice and education. BMC Med Educ. 2006;6:41.*

[2] *Gwozdek AE, Klausner CP, Kerschbaum WE. The utilization of Computer Mediated Communication for case study collaboration. J Dent Hyg. 2008;82(1):8.*

[3] *Jacobs N. Hospital CEO raises awareness through blog. Profiles in Healthcare Communications 23(2):9–11.*

[4] *Poonawalla T, Wagner RF Jr. Assessment of a blog as a medium for dermatology education. Dermatol Online J. 2006;12(1):5.*

[5] *Santoro E. Podcasts, wikis and blogs: the web 2.0 tools for medical and health education. Recent Prog Med. 2007;98(10):484–94.*

[6] *Santoro E. Podcasts, wikis and blogs: the web 2.0 tools for medical and health education. Recent Prog Med. 2007;98(10):484–94.*

[7] Lagu, T., E. Kaufman, D. Asch, and K. Armstrong. 2008. Content of Weblogs Written by Health Professionals. Journal of General Internal Medicine, 23 (10), pages 1642-1646

[8] Kennedy D. Doctor blogs raise concerns about patient privacy. Available at: http://www.npr.org/templates/story/story.php?storyId=88163567. March 13, 2008.

[9] Hillan J. Physician use of patient-centered weblogs and online journals. Clinical Medicine & Research. 2003;1(4):333–5.

[10] Thielst CB. Weblogs: a communication tool. J Healthc Manag. 2007;52 (5):287–9.

[11] Herper M. Best Medical Blogs. Available at: http://www.forbes.com/2003/10/03/cx_mh_1003medblogs.html.

[12] Medical weblog awards: Meet the winners! Available at: http://medgadget.com/archives/2007/01/2006_medical_we.html.

[13] Alvarez M. Is there a blogger in the house? Five great doctor blogs. Available at: http://www.foxnews.com/story/0,2933,246919,00.html.Juanuary , 2007.

[14] Kennedy D. Doctor blogs raise concerns about patient privacy. Available at: http://www.npr.org/templates/story/story.php?storyId=88163567.March,2008;

[15] Painter K. Paging Dr. Blog: Online discourse raises questions. Available at: http://www.usatoday.com/news/health/painter/2007-05-13-doctorblog_N.htm. Accessed May 21, 2008.

[16] Hinduja, S. and J. Patchin. 2008. Personal information of adolescents on the Internet: A quantitative content analysis of MySpace. Journal of Adolescence, 31, pages 125-146

[17] Huffaker, D. and S. Calvert 2005. Gender, identity, and language use in teenage blogs. Journal of Computer-Meditated Communication, 10 (2).

[18] Wright, K. and S. Bell. 2003 Health-related Support Groups on the Internet: Linking Empirical Findings to Social Support and Computer-mediated Communication Theory. Journal of Health Psychology, 8 (1), pages 39-54

[19] Baird, M. 2008 Personal files were accessible for more than three weeks. The Western Star. http://www.thewesternstar.com/index.cfm? sid-=104156&sc=23, retrieved Feb 5, 2009

[20] E. Johnson. 2009. Data hemorrhages in the health-care sector. In Financial Cryptography and Data Security

[21] Long, J. 2008. No Tech Hacking: A Guide to Social Engineering, Dumpster Diving, and Shoulder Surfing. Syngress Press.

[22] Personal Health Information Protection Act. Legislation of Ontario, 2004. http://www.e-laws.-gov.on.ca/html/statutes/english/elaws statutes 04p03 e.htm, accessed Sept. 7, 2008.

[23] Herper M. Best Medical Blogs. Available at: http://www.forbes.com/ 2003/10/03/cx_mh_1003medblogs.html. Accessed May 21, 2008.

[24] Kennedy D. Doctor blogs raise concerns about patient privacy. Available at: http://www.npr.org/templates/story/story.php?storyId=88163567. Accessed May 21, 2008.

[25] Painter K. Paging Dr. Blog: Online discourse raises questions. Available at: http://www.usatoday.com/news/health/painter/2007-05-13-doctorblog_N.htm. Accessed May 21, 2008.

[26] Siegler M. Sounding boards. Confidentiality in medicine-a decrepit concept. N Engl J Med. 1982;307(24):1518–21.

[27] Ubel PA, Zell MM, Miller DJ, Fischer GS, Peters-Stefani D, Arnold RM. Elevator talk: observational study of inappropriate comments in a public space. Am J Med. 1995;99(2):190–4.

[28] Sobel RK. Does laughter make good medicine? N Engl J Med. 2006;354 (11):1114–5.

[29] Wynia MK, Latham SR, Kao AC, Berg JW, Emanuel LL. Medical professionalism in society. N Engl J Med. 1999;341(21):1612–6.

[30] ABIM. Medical professionalism in the new millennium: A physician charter. Available at: http://www.abimfoundation.org/professionalism/charter.shtm. Accessed May 21, 2008.

[31] Tom Buchanan, Adam N. Joinson and Carina Paine;'Looking for medical information on the Internet: self-disclosure, privacy and trust'; Health Info Internet 2007;58:8-9; Royal Society of Medicine Press

32 *Lagu, T., E. Kaufman, D. Asch, and K. Armstrong. 2008. Content of Weblogs Written by Health Professionals. Journal of General Internal Medicine, 23 (10), pages 1642-1646*

33 *Buchanan, T., A. Joinson, C. Paine and U.-D. Reips. 2007. Looking for medical information on the Internet: self-disclosure, privacy and trust Health Information on the Internet, 58, pages 8-9.*

34 *Wright, K. and S. Bell. 2003 Health-related Support Groups on the Internet: Linking Empirical Findings to Social Support and Computer mediated Communication Theory. Journal of Health Psychology, 8 (1), pages 39-54*

35 *Kennedy, D. 2008. Doctor Blogs Raise Concerns About Patient Privacy. National Public Radio. http://www.npr.org/templates/story/story.php?storyId=88163567, retrieved December 15, 2009.*

36 *Silverman, E. 2008. Doctor Blogs Reveal Patient Info & Endorse Products. Pharmalot http://www.pharmalot.com/2008/07/doctor-blogs-reveal-patient-info-endorse-products/, retrieved December 15, 2009.*

37 *Clive Best, Bruno Pouliquen, Ralf Steinberger, Erik Van der Goot, Ken Blackler, Flavio Fuart, Tamara Oellinger, Camelia Ignat;"Automatic Event Extraction for the security domain"; Towards Automatic Event Tracking. ISI 2006: 26-34*

38 *Clive Best, Jakub Piskorski, Bruno Pouliquen, Ralf Steinberger and Hristo Tanev; ;"Automatic Event Extraction for the security domain"; Chapter 2 of the 'Intelligence and Security Informatics' ; ISBN 978-3-540-69207-2; Springer 2008*

39 *Dwyer, C., S. R. Hiltz and K. Passerini. "Trust and privacy concern within social networking sites: A comparison of Facebook and MySpace". Proceedings of the Thirteenth Americas Conference on Information Systems, Keystone, Colorado August 09 – 12 (2007).*

40 *Hinduja, S. and J. Patchin. "Personal information of adolescents on the Internet: A quantitative content analysis of MySpace". Journal of Adolescence, 31, pages 125-146-(2008)*

41. *Hoecht, A. and Trott, P., "Outsourcing, Information Leakage and the Risk of Losing Technology-based Competencies", European Business Review, Vol. 18, No.5, pp. 395-412, (2006)*

42*. Yasuhiro Kirihata, Yoshiki Sameshima," A Web-based System for Prevention of Information Leakage";*

WWW2002, 2002

43 *Vitor Carvalho, William Cohen and Ramnath Balasubramanyan,'Cut Once - A Thunderbird extension for Recipient Prediction and Leak Detection', 2008*
[ECIR-2008] Vitor R. Carvalho, William W. Cohen. Ranking Users for Intelligent Message Addressing. ECIR 2008: 321-333

44.*Lagu T, Kaufman EJ, Asch DA, Armstrong K.,' Content of Weblogs Written by Health Professionals' J Gen Intern Med. 2008 Oct;23(10):1642-6. Epub 2008 Jul 23.*

45 *Ed Silverman ;'Doctor Blogs Reveal Patient Info & Endorse Products'; Pharma Blog » 2008 » July » 23*

46 *Gunnar Carlsson; Facundo M_emoli –" Multiparameter Hierarchical Clustering Methods" March 18,2009*

47 *Hong-Woo Chun, Yoshimasa Tsuruoka, Jin-Dong Kim, Rie Shiba, Naoki Nagata, Teruyoshi Hishiki, Jun-ichi Tsujii (2006). "Extraction of Gene-Disease Relations from Medline Using Domain Dictionaries and Machine Learning". Pacific Symposium on Biocomputing.*

48 *T.C.Rindflesch and L.Tanabe and J.N.Weinstein and L.Hunter (2000). "EDGAR: Extraction of drugs, genes, and relations from the biomedical literature". Proc. Pacific Symposium on Biocomputing. pp. 514--525.*

49 *Amir H. Razavi, Stan Matwin, Diana Inkpen, and Alexandre Kouznetsov, 'Parameterized Contrast in Second Order Soft Co-Occurrences: A Novel Text Representation Technique in Text Mining and Knowledge Extraction', ICDM Workshops 2009*

50 *Stan Matwin, Amir Razavi, Joseph De Koninck, Ray Reza Amini;" Classification of Dreams Using Machine Learning" will be appeared in the proceedings of the Sixth Conference on Prestigious Applications of Intelligent Systems (PAIS 2010)*

Learning to Classify Medical Documents According to Formal and Informal Style

Fadi Abu Sheikha and Diana Inkpen

School of Information Technology and Engineering,
University of Ottawa
{fabus102@uottawa.ca, diana@site.uottawa.ca}

Abstract. This paper discusses an important issue in computational linguistics: classifying sets of medical documents into formal or informal style. This might be important for patient safety. Formal documents are more likely to have been published by medical authorities; therefore, the patients could trust them more than they can trust informal documents. We used machine learning techniques in order to automatically classify documents into formal and informal style. First, we studied the main characteristics of each style in order to train a system that can distinguish between them. Then, we built our data set by collecting documents for both styles, from different sources. After that, we performed pre-processing tasks on the collected documents to extract features that represent the main characteristics of both styles. Finally, we test several classification algorithms, namely Decision Trees, Naïve Bayes, and Support Vector Machines, to choose the classifier that leads to the best classification results.

Keywords: Automatic Text Classification, Medical Documents, Formal Style, Informal Style.

1 Introduction

The need for identifying and interpreting possible differences in linguistic style of medical documents, such as between formal and informal styles, has increased nowadays as more and more people are using the Internet as a main resource for their researches. There are different factors that affect formality, such as words and expressions, as well as syntactical features. Vocabulary choice is perhaps the biggest style marker. Generally speaking, longer words and Latin origin verbs are formal, while phrasal verbs and idioms are informal. There are also many formal/informal style equivalents that can be used in writing.

Formal style is used in most writing and business situations and in speaking with people with which we do not have close relationships. Some characteristics of this style are using long words and passive voice. While Informal style is used in casual conversation, for example, that often happens at home between family members. It is used in writing only when there is a personal or closed relationship, like between friends and family. Some characteristics of this style are using word contractions like "*won't*", abbreviations like "*phone*", and short words.

In this paper we show how to build a model that will help to automatically classifying any medical document into formal or informal style. So, we tested several classification algorithms, namely Decision Trees, Naïve Bayes, and Support Vector Machines in order to choose the classifier that leads to the best classification results.

Automatic classification of medical documents into formal and informal might be important for patient safety, since informal documents are unlikely to be published by medical authorities; therefore, people should not trust informal documents found in Internet.

The rest of the paper is organized as follows: In Section two, we review some existing methods for text classification by style and by genre. Section three addresses the main differences between both styles. In Section four, we discuss how we collect our data set that will be used to train our model. Section five presents our approach for extracting the features to build our model. In Sections six, we describe the classification algorithms that we used to train our model. Section 7 addresses the result and the evaluation methods for our model. In Sections 8 we discuss the results that we obtained. Finally, Section 9 concludes the paper and discuses the future work.

2 Related Work

There is little research on automatic text classification according to formal and informal style. For instance, Heylighen and Dewaele (1999) proposed a method to determine the degree of formality for any text using a special formula. This formula is the F-score measurement which is based on the frequencies of different word classes (noun, verbs, adverbs, etc.) in the corpus. The texts with high F-score are considered formal, while the ones with low F-score are considered informal. In our work, we want to build a model based on main characteristics of the two styles, rather than based on the frequency of word classes.

Moreover, Dempsey, McCarthy & McNamara (2007) propose that phrasal verbs can be used as a text genre identifier. Their results indicate that phrasal verbs significantly distinguish between both the spoken/written and the formal/informal dimensions. Their experiments are performed on the frequency of occurrence of phrasal verbs in spoken versus written text and in formal versus informal texts.

In addition, there is some work on automatic text classification by genre. Of course, there is a lot of research on classifying texts by their topic, but this does not apply in our case, since the texts can have different styles and be about the same topic. Similarly the texts can be about different topics and have the same style.

3 Learning Formal and Informal Style

In this section, we explain the main characteristics for formal versus informal style. We also show a sample of ready-made list of words for both styles, which we collected from different sources; this will help to understand the difference between the two styles.

3.1 Characteristics of Formal versus Informal Style

We studied and summarized the main characteristics of formal style versus informal style from Dumaine and Healey (2003), Obrecht and Ferris (2005), and Akmajian et al (2001) to:

- Be able to distinguish between both styles.
- Identify each style from texts.
- Build the features based on those characteristics.
- Predict a class for new text documents.

Here we explain the characteristics of each style and provide examples:

Main Characteristics of Informal Style Text

1. It uses a personal style, using the first and second person (*I*, *you*) and the active voice (e.g., *I have noticed that...*).
2. It uses short simple words and sentences.
3. It uses Contractions (e.g., *won't*) and abbreviations (e.g., *TV*).
4. It uses phrasal verbs (Anglo Saxon words) within the text (e.g., *find out*).
5. The words that express rapport and familiarity are often used in speech, such as *brother, buddy, and man*.
6. It is more used in everyday speech than in writing.
7. It uses a subjective style, expressing opinions and feelings (e.g., *pretty, I feel*).
8. It uses vague expressions, it uses personal vocabulary and colloquial (slang words are accepted in spoken not in written text (e.g., *wanna* = *want to*)).

Main Characteristics of Formal Style

1. It uses an impersonal style, using the third person (*it, he,* and *she*) and often the passive voice (e.g., *It has been noticed that*….).
2. It uses complex words and sentences to express complex points.
3. It does not use contractions or abbreviations.
4. It uses appropriate and clear expressions, precise education, business, and technical vocabulary (Latin origin).
5. It uses polite words and formulas like (e.g., *Please, Thank you, Madam, Sir*)
6. It is more commonly used in writing than in speech.
7. It uses an objective style, using facts and references to support an argument.
8. It does not use vague expressions and slang words.

3.2 Formal versus Informal list of words

We collected informal/formal words, phrases, and expressions from different sources manually, also we extracted automatically more words from annotated text documents; such lists were very useful as two of the features in our model. In Table 1, we show an example of this list.

Table1. An example of formal versus informal list of words

Informal	**Formal**
about	approximately
and	in addition
anybody	anyone
ask for	request
boss	employer
but	however
buy	purchase
end	finish
enough	sufficient
get	obtain
go up	increase
have to	must

4 Data Set

The size of the data set that we collected is 1980 documents: 990 characterize informal text and 990 characterize formal text.

Informal Texts

We chose 990 texts that characterize the informal style (Yu-shan & Yun-Hsuan 2005) from Medical newsgroups collection, this corpus called 20 Newsgroups[1] contains 20 topics, and each topic has 1000 texts. These texts characterize informal style. We use one of these topics which are medical texts. We excluded 10 documents which have less than two words.

Formal Texts

We chose randomly 990 texts that characterize the formal style from medical abstracts collection. This collection contains 23 cardiovascular diseases categories (Joachims, 1997).

[1] http://kdd.ics.uci.edu/databases/20newsgroups/20newsgroups.html

5 Features

We built features that characterize formal and informal texts, based on the above analysis in section 3. We hypostasized that these features might be a good indicator to differentiate between both styles. We applied several statistical methods in order to extract the values of these features for each text in our dataset. Some of the features required us to parse each text. We parsed all the documents with the Connexor parser[2], which helps to produce high-quality results for our model.

The features that we extracted are as follows:

1. **Formal words list**: This feature is based on the formal list that we had mentioned in section 3.2. The value of this feature is based on its frequency in each text normalized by the length of the text for each document.
2. **Informal words list:** This feature is based on the informal list. The value of this feature is based on its frequency in each text normalized by the length of the text for each document.
3. **Formal pronouns:** This feature characterizes formal texts. In the parse trees returned by the Connexor parser, we counted the frequency of impersonal pronouns, and we normalized by the length of the text for each document.
4. **Informal pronouns:** This feature characterizes informal texts. In the parse trees returned by the Connexor parser, we counted how many times the text has personal pronouns normalized by the length of the text for each document.
5. **Contractions:** This feature characterizes informal texts. We counted the contractions words normalized by the length of the text for each document.
6. **Abbreviations:** This feature characterizes informal texts. We counted the abbreviations normalized by the length of the text for each document.
7. **Passive voice:** This feature characterizes formal texts. In the parse trees returned by the Connexor parser, we counted how many times the text has a passive voice normalized by the length of the text for each document.
8. **Active voice:** This feature characterizes informal texts. In the parse trees returned by the Connexor parser, we counted how many times the text has an active voice normalized by the length of the text for each document.
9. **Phrasal verbs:** This feature characterizes informal texts. In the parse trees returned by the Connexor parser, we counted how many times the text has phrasal verbs normalized by the length of the text for each document.
10. **Word length's average:** This feature characterizes formal texts, if the value is large (complex words), and it characterizes informal texts if the value is small (simple words). We calculated the average for the words for each document.
11. **Type Tokens Ratio (TTR):** This feature refers to how many distinct words are in a text comparing to the total number of words in the text.

We used a parser to obtain some of the features. For most of them, a part-of-speech tagger would have been enough, but for some features the extra information provided by the parser was needed, for example for active/passive voice and for phrasal verb.

[2] http://www.connexor.com

6 Classification Algorithms

We used WEKA[3] (Witten & Frank 2005), a collection of machine learning algorithms for data mining tasks. The algorithms can either be applied directly to a certain dataset or called from Java code. WEKA contains tools for data pre-processing, classification, regression, clustering, association rules, and visualization. It is also well-suited for developing new machine learning schemes.

We chose three machine learning algorithms (Witten & Frank 2005): Decision Trees because it allows human interpretation of what is learnt, Naïve Bayes because it is known to work well with text, and Support Vector Machines (SVM) because it is known to achieve high performance. Table 2 shows the classification result for the three classifiers, by 10-fold cross-validation on our data set.

7 Results and Evaluation

As we mentioned in section 6, we trained three classifiers: Decision Tree, Naïve Bayes, and SVM. . The Experiments were run using a 10-fold cross validation test. Results are shown in Table 2 for all three classifiers. The standard evaluation metric of F-Measure, the weighted harmonic mean of precision and recall was calculated. The Results show that SVM was the best classifier for our model that has achieved best performance. In Table 3, we show the detailed F-measure per class of SVM algorithm. Finally, we examined all the features by performing attribute selection using InfoGain attribute selection (InfoGainAttributeEval) from Weka. We tried to remove the weakest features but we discovered that this decreased the accuracy for the three algorithms. So, we decided to keep all the features in our model, as all features are important to achieve good performance. Table 4, shows each attribute with its weight according to the InfoGain attribute selection, ranked in descending order from the strongest features to the weakest features. The most useful feature was the average word length.

Table2. Classification results of SVM, Decision Trees, and Naïve Bayes classifiers.

Machine Learning Algorithm	**F-measure (Weighted Avg.)**
Support Victor Machine (SMO)	0.977
Decision Trees (J48)	0.972
Naïve Bayes (NB)	0.965

[3] http://www.cs.waikato.ac.nz/ml/weka/

Table3. Detailed accuracy for both classes of SVM

Class	Precision	Recall	F-Measure
Informal	0.991	0.963	0.976
Formal	0.964	0.991	0.977
Weighted Avg.	0.977	0.977	0.9

Table4. Our model’s features with their InfoGain scores

Attributes	Weight
Word length’s average	0.745
Active voice	0.5719
Informal pronouns	0.5636
Contractions	0.4571
Passive voice	0.2192
Informal list	0.1913
Type Tokens Ratio (TTR)	0.1598
Formal pronouns	0.0913
Formal list	0.0815
Phrasal verbs	0.0748
Abbreviations	0.0168

8 Discussion

Our experiments show that it is possible to classify any Medical text according to formal and informal style. We achieved reliable accuracies for all three classifiers, especially on SVM. This indicates that we selected high quality features to include in our model. This model can generate good results whether it is applied on a single topic or on different topics.

9 Conclusion and Future Work

In this paper we have discussed one approach to classify medical documents according to formal and informal style. In doing so we presented the main characteristics of both styles. From these characteristics we derived the features of our model. The learning process was successful and the classifiers were able to predict the classes of new texts with high accuracy.

Our immediate future work will be on extracting more formal and informal lists which should increase the accuracy of the classifiers. We will also experiment with adding more features such as sentence length feature in order to obtain a classifier with close to 100% accuracy.

Acknowledgements

Our research is supported by the Natural Sciences and Engineering Research Council of Canada (NSERC), the University of Ottawa. We thank Soumaya Chaffar and Alistair Kennedy for helpful comments on our draft.

References

Akmajian, Adrian; Demers, Richard A.; Farmer, Ann K.; & Harnish, Robert M. (2001). “Linguistics: an introduction to language and communication”, (pp. 287-291), 5th Edition, MIT Press, Cambridge (MA).

Dempsey, K.B., McCarthy, P.M., & McNamara, D.S. (2007). “Using phrasal verbs as an index to distinguish text genres”. In D. Wilson and G. Sutcliffe (Eds.), Proceedings of the twentieth International Florida Artificial Intelligence Research Society Conference (pp. 217-222). Menlo Park, California: The AAAI Press.

Dumaine, Deborah & Healey, Elisabeth C (2003). “Instant-Answer Guide To Business Writing: An A-Z Source For Today's Business Writer,”(pp. 153-156), 2003 Edition, Writers Club Press, Lincoln.

Heylinghen, Francis & Dewaele, Jean-Marc. 1999 “Formality of language: definition and measurement”. Internal Report, Center "Leo Apostel", Free University of Brussels.

Ian H. Witten; Eibe Frank (2005). "Data Mining: Practical machine learning tools and techniques”. 2nd Edition, Morgan Kaufmann, San Francisco.

Obrecht, Fred & Ferris, Boak (2005). “How to Prepare for the California State University Writing Proficiency Exams”(pp. 173), 3rd Edition, Barron’s Educational Series Inc., New York.

Thorsten Joachims (1997), “Text Categorization with Support Vector Machines: Learning with Many Relevant Features”. LS8-Report 23, Universitat Dortmund, LS VIII-Report.

Yu-shan, Chang & Yun-Hsuan, Sung (2005). “Applying Name Entity Recognition to Informal Text”, Ling 237 Final Projects.

Risk Analytic Framework for DLP Systems

Ahmed Al-Faresi, Ahmed Alazzawe, Anis Alazzawe, and Duminda Wijesekera

George Mason University, Department of Computer Science,
4400 University Drive, Fairfax, Virginia 22030
{aalfares,aalazza1,aalazzaw,dwijesek}@gmu.edu

Abstract. Data leakage protection systems (DLPs) have emerged as powerful tools to detect and secure sensitive information. In health care sensitive data is present in various formats. This variation ranges from structured data like in databases to unstructured data like in e-mails, e-documents and forum posts.The privacy risk associated with such diverse health information formats varies accordingly. Quantifying and classifying privacy risk associated with the various forms of data would enable privacy mangers to better secure the confidentiality of the data.This paper presents a risk analysis framework to be integrated with DLPs. We use a Bayesian network model to aid DLPs in quantifying the associated risk with attributes in unstructured data sets. This in turn, enables categorizing the different health data forms according to a risk level. We also propose an architecture that integrates the risk anlysis method with DLPs.

Keywords: Privacy - Re-Idenfication risk - DLP - HIPAA - PHI

1 Introduction

The surge in demand for sharing and processing health information, made the task of protecting patient privacy extremely immense. At the core of protecting patient privacy lies our ability in securing the confidentiality of patient health information. Protected health information (PHI) or ePHI (i.e. if in electronic form), under the US Health Insurance Portability and Accountability Act (HIPAA), is any information about health status, provision of health care, or payment for health care that can be linked to a specific individual. Usually PHI refers to explicit patient identifiers like name, address, birth date, medical record numbers etc. In this paper we will use PHI and ePHI to refer to explicit identifiers as stated by HIPAA.

ePHI is often stored on multiple devices and shared in diverse formats. Data leakage protection systems (DLPs) have been deployed to locate ePHI in formats other than the medical records like e-documents, nurse notes, emails, forum posts and end point devices, which we will refer to as unstructured health information (UHI). However many such commercial DLPs [1] [7] [11]. focus on de-identifying data by detecting and masking/encrypting explicit health identifiers (i.e ePHI), but fail to account for the potential privacy leakage risk, associated with other attributes in UHI. These attributes are commonly called quasi–identifiers (QIs) and may include attributes such as gender, ethnicity, and profession. An adversary can match the QIs in UHI against auxiliary sources such as public databases, public registries, and social networks to potentially re-identify the individual to whom the UHI pertains to. Moreover it may beneficial for privacy advisors to manage UHI, detected by a DLPs, if such UHI is labeled with quantifiable risk values.The National Institute of Standards and Technology (NIST) has published a draft guide for protecting personally identifiable information (PII) based on the level of sensitivity [6]. The guide suggests that not all PII is to be treated the same, which in turn, suggests that data should be classified and tagged with quantifiable risk values. DLPs should incorporate risk measures in their classification of QIs and ePHI, as best practices or to comply with such guidelines if they become standard.

In this paper we describe a method to quantify the risk of QIs that would typically be used in a healthcare related emails or forum posts. We also propose an architecture to integrate this method within the DLP. The results of this approach will assist DLPs to detect high risk QIs or ePHI and categorize documents into appropriate risk values accordingly. Furthermore quantifying the risk of QIs will aid in designing privacy-training tools that would enable employees to write better e-mails or posts, by being aware of risky QI they need to avoid.

2 Related Work

In this section we describe the methods of de-identification, re-identficaiton and examine previous work done on risk analysis for re-identification.

2.1 Identification

Under HIPAA privacy rule there are two methods for de-identification

(1) A statistician, with appropriate expertise in applying generally accepted statistical and scientific principles and methods for making information not individually identifiable, determines that the risk is very small that the information could be used either by itself, or in combination with other available information by anticipated recipients to identify an Individual.

(2) Removal of specified list of eighteen categories of possible identifiers as they pertain to the Individual or to his/her relatives, employers or household members. This method is commonly known as the ”Safe Harbor” rule.

A considerable amount of research has been done in the automatic de-identification of ePHI by applying the ”Safe Harbor” rule. Most of these initiatives such as [12] [10][9][13] utilized named entity recognition (NER) techniques (i.e. also known as entity identification and entity extraction) to de-identify ePHI in UHI data sets. Current DLPs deployed in healthcare environments rely on such technology for ePHI detection.In this work we do not discuss any techniques to detect QIs but we make the assumption that automatic detection of QIs is possible by extension on the work done on ePHI in that regard.

2.2 Re-identification

Re-identification is the process where seemingly none identifying information is mapped to explicit identifying information. The mapping can be achieved by multiple techniques such as inference and record linkage [8]. Record linkage is a set of techniques used to match QIs in de-identified data sets with auxiliary sources to obtain indentifying information. There are commonly three re-identification attacks found in literature [4] namely:

1- Prosecutor: An adversary attempts to re-identify a specific person in a de-identified data set.
2- Journalist: An adversary knows that all the individuals represented in the de-identified dataset exist in a larger public database.
3- Marketer: An adversary attempts to re-identify as many individuals as possible in the de-identified dataset.

In this work we are concerned with an attack that re-identifies as many individuals as possible in UHI datasets (i.e. marketer attack). As previously mentioned UHI can be an email a forum post or an e-document. We make the assumption that each UHI dataset pertains to a single individual in contrast to the micro data model [14] where multiple individuals are represented in the same dataset.

2.3 Re-identification risk analysis

Benitez et al [2] provides an approach for estimating the likelihood that de-identified information can be re-identified in the context of data sharing policies associated with the HIPAA privacy rule. In their work a risk metric was proposed to evaluate the expected number of re-identifications of the released data set. The work shows de-identified data under the safe harbor rule tested with the following QIs (Year of Birth, Gender, Race) over voters lists. The re-identification risks measured ranged from from 0-1% over 50 states. The test was repeated with the addition of (County and date of birth) and the risk estimates ranged from 10-60% over 50 states. The study reveals that different organizations can be vulnerable to re-identification. Our approach aims at addressing this shortcoming by proposing that DLPs perform risk analysis to quantify the risk associated with de-identified UHI.

El Emam et al [5] evaluated the re-identification risks due to record linkage by following common de-identification heuristics. The study used two identification data sets namely the list of all physicians and the list of all lawyers registered in Ontario. The data set, the sample size, and the QI combinations were varied and evaluated. The QIs of (region, gender, and year of birth) were found to be low risk more than 50% of the time across both data sets. The combination of gender and region was also found to be low risk more than 50% of the time. This study used Data Intrusion Simulation (DIS) sampling technique [3]. The idea here is to simulate a sample of the identification data set and the released data set. This relieves the need to construct an identification database to estimate the probability of exposure. The risk estimate used was based on the probability of finding a unique match of a record in the identification database with a record in the released data set.

Both studies have considered a marketer attack and are based on the probability of re-identification of released micro data sets. In our approach we also consider the scenario of a marketers attack but there is no released dataset known in priori, instead the DLPs will detect varying combinations of QIs in UHI datasets as they are being created. This will require continuous risk assessment of UHI data sets. We adopt a Bayesian network model to estimate the risk associated with the different combination of QIs as they are detected by DLPs. The advantage of this risk estimate, is that it allows us to model more auxiliary data sources with variable reliability. This is useful in cases where some QIs are not found in more comprehensive databases like voters lists. The other advantage is the intended record in each UHI is matched against an axillary source and not a record from a sub sample, this should give a more accurate probability measure.

3 Method

We use a Bayesian network model to estimate the risk of re-identification and illustrate the concept by giving two motivating examples.

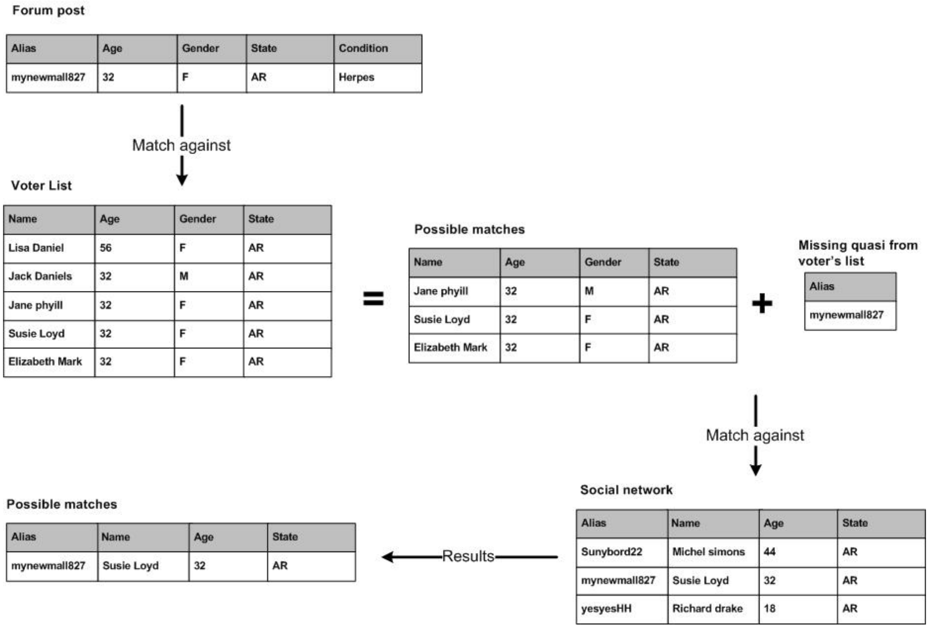
3.1 The "e-mail" example

Imagine a scenario were a nurse in a health institution sends an unencrypted email containing patient ePHI to some recipient. Assume a DLPs was successful in detecting and removing ePHI from the email. The recipient of the email may wish to act as an adversary and re-identify the individual from the QI in the context of the email. The email alias of the sender also acts as an additional QI that enables the recipient to infer additional information. To alleviate this threat a DLPs, would have to be connected to a "white list" of allowed recipients, such that an email is sent only if its associated risk is below the threshold and the recipient email falls within the "white list".

3.2 The "Forum post" example

In this example we consider a re-identification attack using a forum post entry that passed through a DLPs successfully. Typically a health forum post will include a user alias among other QIs to facilitate correspondence among forum members. An example of a re-identification scenario that involves a forum post is depicted in figure 1.

In this example an adversary queries the QIs of (Age,Gender,State) against a voter's list in the same state.The result is a 1/3 chance of a match. Since Alias is QI that obviously does not exist in such a database it was combined with the three other possible matches from the voter's list. This created a sort of new target record for re-identification. The target record was matched against a



Forum post

Alias	Age	Gender	State	Condition
mynewmall827	32	F	AR	Herpes

Voter List

Name	Age	Gender	State
Lisa Daniel	56	F	AR
Jack Daniels	32	M	AR
Jane phyill	32	F	AR
Susie Loyd	32	F	AR
Elizabeth Mark	32	F	AR

Possible matches

Name	Age	Gender	State
Jane phyill	32	M	AR
Susie Loyd	32	F	AR
Elizabeth Mark	32	F	AR

Missing quasi from voter's list

Alias
mynewmall827

Social network

Alias	Name	Age	State
Sunybord22	Michel simons	44	AR
mynewmall827	Susie Loyd	32	AR
yesyesHH	Richard drake	18	AR

Possible matches

Alias	Name	Age	State
mynewmall827	Susie Loyd	32	AR

Fig. 1. Record linkage using forum posts

social database which was unique on the Alias QI and gave one possible match. The probability may seem to be a 100% match but that is not the case. The reason being is that the social network has two variables affecting its reliability. The total percentage of the population in the state of VA present in the database and the confidence level in the integrity of data. The assumption being made here also is that the Voter's list represents the total population. Nevertheless using an additional auxiliary source like a social network in this example does improve the probability of finding a match. It is worth noting that certain QIs in auxiliary data sets could be used to infer non-existing QIs in that same data set. For example if one is able to retrieve a data set from a social network containing a name and age, then one could infer gender from the name. Thus one can produce additional QIs using inference.

3.3 Risk estimation using a single auxiliary source

To quantify this risk in a UHI data set, we calculate the probability of re-identifying an individual from the set of QIs present in the UHI. In this calculation we assume that an attacker attempts re-identification by accessing one data source. Generally the probability of re-identification using a given auxiliary dataset and a set of QIs is given by the following:

$$Risk = P_{aux}(\text{Re-Identification}|q_1...q_n) \ . \tag{1}$$

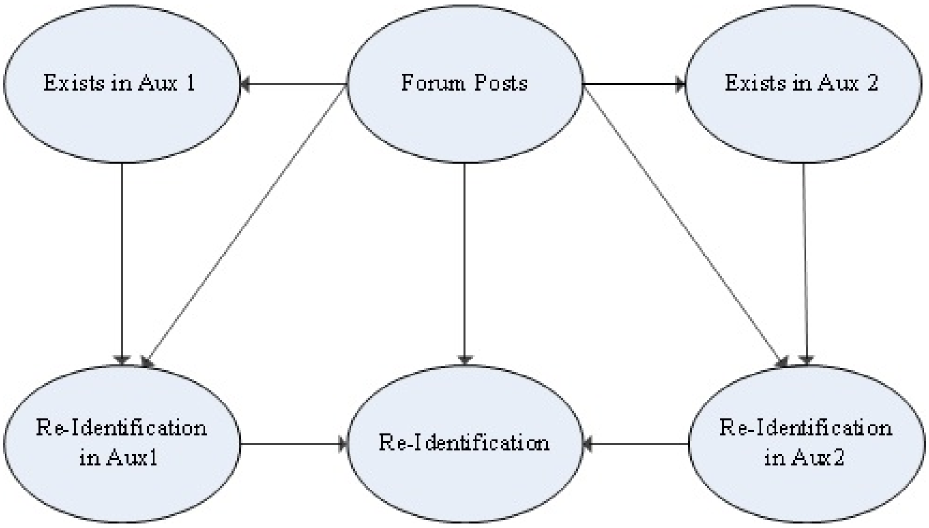


Fig. 2. Model of linkage

The P_{aux} in formula 1 cannot be calculated directly. Rather the risk is composed of two components which we will call R_1 and R_2. R_1 is the probability of identifying the subject in the UHI given a set of QIs and the fact that the subject exists is in the *aux* data source. R_2 is the probability that the subject exists in the auxiliary data source. So we define risk as $Risk = R_1 * R_2$.

$$R_1 = P(\text{Re-Identification}|q_1...q_n, Subject\ exists\ in\ aux) = \frac{1}{\widehat{E}} . \qquad (2)$$

In equation (2), $\widehat{E}$ represents the equivalence class, which is the set of QI values in the UHI and auxiliary data source that match. Bad entries in the auxiliary data source, the UHI containing inaccurate QIs, or there really is no chance the subject is in the dataset in which case the R_2 is zero.

$$\begin{aligned} R_2 &= P(Subject\ in\ aux|q_1...q_n) . \\ R_2 &= \frac{P(q_1...q_2|Subject\ in\ aux) * P(Subject\ in\ aux)}{P(q_1...q_n)} \end{aligned} \qquad (3)$$

Here R_2 is the probability that the Subject is in the auxiliary dataset given a set of QIs extracted from UHI data set. Using Bayes' theorem we can calculate this using the probability of finding QIs given a match in the auxiliary data source multiplied by the probability of finding a subject in the same auxiliary data source.

3.4 Risk estimation in multiple auxiliary sources using probabilistic modeling

We appeal to bayesian networks as a way of modeling and identifying the risk in a multi-aux environment. The idea is to explicitly identify the dependencies between the different components of the risk factor. So in a two aux environment we discussed previously in figure 1, we need to capture the risk of re-identification in the voter's list as well as in the social network. We then capture the risk of combining the two together. Doing this we end up with a graphical model in figure 2. This model can answer many questions, but the one that is pertinent to the DLP is the query $P(\text{Total Re-Identification}|QIs)$. Back to our example, this would give the risk of re-identification using both the voter list and the social network data given a set of QIs extracted from an email or forum post.

For each component of the network we build a Conditional Probability Table (CPT). For each node, we are only concerned about the state of the parent node, because the model states that the parent node is the only influence on the probability of the current node. So the constructed CPT will express the probabilities of derived child node from the parent nodes conditions.

In general using this network we can describe the full joint distribution as

$$P(x_1, ..., x_n) = \prod_{1}^{n} P(x_i|parents(X_i)) \tag{4}$$

We can also compute the marginals as

$$P(x_1, ..., x_m) = \sum_{x_{m+1},...,x_n} P(x_1, ..., x_n) \tag{5}$$

And given some evidence such as QIs from a forum post

$$P(x_1, ..., x_m|evidence) = \sum_{x_{m+1},...,x_n} P(x_1, ..., x_n|evidence) \tag{6}$$

From equation 4, 5, and 6 we can easily calculate $P(\text{Total Re-Identification}|QIs)$. We then can use the equations 2 and 3 for calculating the existence and re-identification nodes.

4 Architecture

We propose an architecture model to integrate risk analysis in a DLP as depicted in figure 3. The model makes use of our proposed risk analysis method. The architecture is compromised of four modules namely the parser, classifier, risk analysis engine and the decision control unit. The parser parses the unstructured document and sends it to the classifier. The classifier detects any QIs present within context. The extracted QIs are sent to both the risk analysis engine and the decision control unit. If the online mode is active, the risk analysis engine will dynamically check for the risk associated with the QIs values being supplied. The risk values will be relayed to the decision control unit for a decision. If the batch mode is used, then the values will be ignored and the decision control unit will base a risk analysis decision on the programmed heuristics.

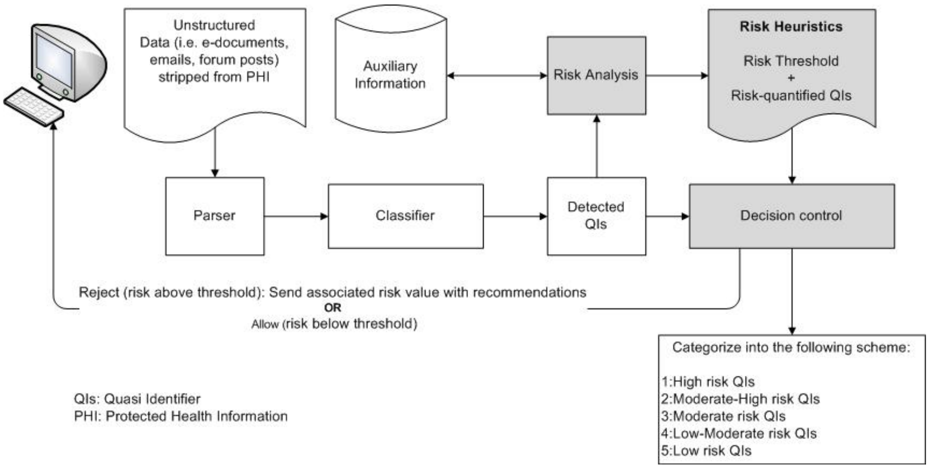

Unstructured Data (i.e. e-documents, emails, forum posts) stripped from PHI
Auxiliary Information
Risk Analysis
Risk Heuristics
Risk Threshold + Risk-quantified QIs
Parser
Classifier
Detected QIs
Decision control
Reject (risk above threshold): Send associated risk value with recommendations
OR
Allow (risk below threshold)
Categorize into the following scheme:
1:High risk QIs
2:Moderate-High risk QIs
3:Moderate risk QIs
4:Low-Moderate risk QIs
5:Low risk QIs
QIs: Quasi Identifier
PHI: Protected Health Information

Fig. 3. Proposed architecture of classifying e-documents with risk in a DLP

5 Conclusion

We have described an analysis method to quantify risk with releasing the QIs in an e-document. We also proposed an architecture to integrate this method into a DLP. Future work would test this method with sample data taken from the referenced forum posts.

Risk of re-identification of unstructured data like in emails depends heavily on context. Assuming DLPs are able to detect explicit identifiers within the context of a document accurately, it still has to determine if QIs present could contribute a significant risk of re-identification.

References

1. Code green networks | complete data loss prevention, `http://www.codegreennetworks.com/products/index.htm`
2. Benitez, K., Malin, B.: Evaluating re-identification risks with respect to the HIPAA privacy rule. Journal of the American Medical Informatics Association 17(2), 169–177 (2010), `http://jamia.bmj.com/content/17/2/169.abstract`

3. Elliot, M.: DIS: a new approach to the measurement of statistical disclosure risk. Risk Management 2(4), 39–48 (2000)
4. Emam, K.E., Dankar, F.K.: Protecting privacy using k-anonymity. Journal of the American Medical Informatics Association 15(5), 627 (2008)
5. Emam, K.E., Jabbouri, S., Sams, S., Drouet, Y., Power, M.: Evaluating common de-identification heuristics for personal health information. Journal of Medical Internet Research 8(4), e28 (2006), `http://www.jmir.org/2006/4/e28/HTML`
6. McCallister, E., Grance, T., Kent, K., of Standards, N.I., (U.S.), T.: Guide to protecting the confidentiality of Personally Identifiable Information (PII) (draft) [electronic resource] : recommendations of the National Institute of Standards and Technology / Erika McCallister, Tim Grance, Karen Scarfone. U.S. Dept. of Commerce, National Institute of Standards and Technology, Gaithersburg, MD :, draft. edn. (2009)
7. A comprehensive approach to regulatory compliance, `www.rsa.com/products/EDS/sb/DLPRC_SB_1107-lowres.pdf`
8. Sweeney, L.: k-anonymity: a model for protecting privacy. Int. J. Uncertain. Fuzziness Knowl.-Based Syst. 10(5), 557–570 (2002)
9. Szarvas, G., Farkas, R., Busa-Fekete, R.: State-of-the-art Anonymization of Medical Records Using an Iterative Machine Learning Framework. Journal of the American Medical Informatics Association 14(5), 574–580 (2007), `http://jamia.bmj.com/content/14/5/574.abstract`
10. Uzuner, O., Luo, Y., Szolovits, P.: Evaluating the state-of-the-art in automatic de-identification. J Am Med Inform Assoc 14(5), 550–63 (2007), `http://www.biomedsearch.com/nih/Evaluating-state-art-in-automatic/17600094.html`
11. Vericept data loss prevention, `http://www.mblast.com/files/companies/93505/Vericept\%20Data\%20Loss\%20Prevention\%20Brochure.pdf`
12. Wang, Y., Liu, H., Geng, L., Keays, M.S., You, Y.: Automatic detecting documents containing personal health information. In: AIME ’09: Proceedings of the 12th Conference on Artificial Intelligence in Medicine. pp. 335–344. Springer-Verlag, Berlin, Heidelberg (2009)
13. Wellner, B., Huyck, M., Mardis, S., Aberdeen, J., Morgan, A., Peshkin, L., Yeh, A., Hitzeman, J., Hirschman, L.: Rapidly Retargetable Approaches to De-identification in Medical Records. Journal of the American Medical Informatics Association 14(5), 564–573 (2007), `http://jamia.bmj.com/content/14/5/564.abstract`
14. Winkler, W.E.: Re-identification methods for masked microdata. In: Privacy in Statistical Databases. pp. 216–230 (2004)