HIDE: Heterogeneous Information DE-identification*

James Gardner, Li Xiong, Kanwei Li, James J. Lu
Department of Mathematics and Computer Science
Emory University
Atlanta, GA
{jgardn3,lxiong,kli,jlu}@emory.edu

ABSTRACT

While there is an increasing need to share data that may contain personal information, such data sharing must preserve individual privacy without disclosing any identifiable information. A considerable amount of research in the data privacy community has been devoted to formalizing the notion of identifiability with many techniques for anonymization, but is focused exclusively on structured data. On the other hand, efforts on de-identifying medical text documents in the medical informatics community are highly specialized for specific document types or a subset of identifiers. In addition, they rely on simple identifier removal or grouping techniques and do not take advantage of the research developments in the data privacy community. We developed an integrated system, HIDE, for Heterogeneous Information DE-identification including structured and unstructured data utilizing existing anonymization techniques. We demonstrate a prototype of our system and show the effectiveness of our approach through a set of real data augmented with synthesized data.

1. INTRODUCTION

Current information technology enables many organizations to collect, store, and use various types of information about individuals. Government and organizations increasingly recognize the critical value and enormous opportunities in sharing such a wealth of information. However, such data sharing has been stymied by restrictions and concerns about the privacy of individuals. For example, the Shared Pathology Informatics Network (SPIN)[1] is an initiative by The National Cancer Institute for researchers throughout the U.S. for sharing pathology-based data sets annotated with clinical information to discover and validate new diagnostic tests and therapies, and ultimately to improve patient care. However, individually identifiable information is protected under the Health Insurance Portability and Accountability Act (HIPAA)[2]. It is necessary for each institution to de-identify the data before having it accessible.

The problem of data anonymization has drawn a large amount of attention in recent years in the data privacy community. The main objective is for a *data custodian* to generate an anonymized view of the data that does not contain individually identifiable information so that it can be released to a shared network or individual institutions and researchers (*data recipients*). Most of the work in this area has been focused on formalizing the notion of privacy through *identifiability*, and on developing computational approaches that guarantees sufficient privacy protection of a dataset. A few *principles* have been proposed that serve as criteria for judging whether a published dataset provides sufficient privacy protection including k-anonymity and later principles that remedy its problems (e.g. [18, 12, 11, 20, 15]). A large body of work contributes to transforming a dataset to meet a privacy principle (dominantly k-anonymity) using techniques such as generalization, suppression (removal), permutation and swapping of certain data values while minimizing certain quality metrics (e.g. [19, 2, 4, 3, 10, 21]).

While the current research on privacy preserving data publishing has made great progress, its practical utilization lags behind. An overarching complexity of practical applications, but often overlooked in data privacy research, is data heterogeneity. Personal data (such as medical information) resides in both structured (such as discrete lab results) and unstructured forms (such as lab and pathology reports). Unfortunately, the bulk of data privacy research including aforementioned privacy principles and data anonymization techniques focus exclusively on structured data. There are some efforts on de-identifying medical text documents in the medical informatics community (e.g. [17, 6, 16, 1]), however, most of them are highly specialized for specific document types or a subset of identifiers. In addition, they rely on simple identifier removal or grouping techniques and do not take advantage of the research developments in the data privacy community.

Contributions. We developed an integrated system, HIDE [5], for Heterogeneous Information DE-identification including structured and unstructured data. The specific components and contributions of our framework are as follows.

*The research is partially supported by Emory URC and ITSC grants.

[1]Shared Pathology Informatics Network (SPIN). http://spin.nci.nih.gov

[2]Health Insurance Portability and Accountability Act (HIPAA). http://www.hhs.gov/ocr/hipaa/. State law or institutional policy may differ from the HIPAA standard and should be considered as well.

EDBT '09 Saint-Petersburg, Russia

- *Data Linking.* In order to preserve privacy for individuals and apply advanced anonymization techniques in the heterogeneous[3] data space, we propose a *person-centric identifier view* of the data with relevant information mapped or linked to each individual.
- *Identifying and Sensitive Information Extraction.* We leverage the latest Named Entity Extraction techniques from natural language processing to effectively extract identifying and sensitive information from the unstructured data.
- *Anonymization.* We adopt the latest data anonymization techniques to perform data suppression and generalization on the identifier view to anonymize the data based on a given privacy principle.

While we utilize off-the-shelf techniques for some of these components, the main contribution of our system is that it bridges the research on data privacy and text management, and provides an integrated framework that allows the anonymization of heterogeneous data for practical applications. We demonstrate a prototype of our system through a set of real-world data augmented with synthesized data and show the effectiveness of our approach.

2. HIDE FRAMEWORK

We first present an overview of our framework, followed by a discussion on the key components.

2.1 Overview

Our framework consists of a number of key components that integrate de-identification for a heterogeneous data space. Figure 1 presents an illustration of the framework. We present an overview below and give more details on the important components in subsequent subsections.

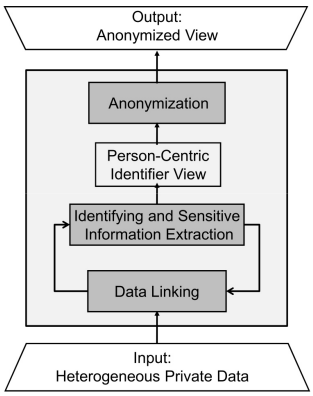

Figure 1: Integrated Framework Overview

In relational data, we assume each tuple corresponds to an individual entity. This mapping is not present in heterogeneous data repository. For example, one patient may have multiple pathology and lab reports prepared at different times. In order to preserve privacy for individuals at a sufficient level, the *data linking* component links relevant attributes (structured attributes or extracted attributes from unstructured data) to each individual entity and produces a person-centric representation of the data. Linking attributes to entities is itself a challenging problem. Our current system uses simple heuristics, e.g. the attributes extracted from a single report are linked to a single patient. It is on our future research agenda to investigate relationship extraction techniques and generalize the approach. We also use a probabilistic record linkage tool we recently developed [7] to resolve potential attribute conflicts and semantic variations in linking records.

While some identifying attributes can be clearly defined in structured data, an extensive set of identifying information is often hidden or have multiple and different references in the text. The *identifying and sensitive information extraction* component extracts the identifying information including identifiers as well as sensitive attributes from unstructured data. Note that this is a much broader set of information to be extracted than existing medical text de-identification systems that typically focus on the set or a subset of HIPAA identifiers.

A novel aspect of our framework is that the data linking component and information extraction component form a feedback loop and are carried out in an iterative manner. Once attributes are extracted from unstructured information, they are linked or added to existing or new entities. Once the data are linked, the linked or structured information will in turn be utilized in the extraction component in the next iteration. The output will be an *identifier view* consisting of identifiers, quasi-identifiers, and sensitive attributes. This notion of identifier view will allow application of current anonymization algorithms that are otherwise not applicable to unstructured data.

Given an identifier view of the integrated heterogeneous data, the *anonymization* component anonymizes the data using generalization and suppression (removal) techniques with different privacy models. Finally, using the generalized values in the anonymized identifier view, we can remove or replace the identifiers in the original data.

2.2 Attribute Extraction

Extracting atomic identifying and sensitive attributes (such as name, address, and disease name) from unstructured data can be seen as an application of named entity recognition (NER) problem [14]. NER systems can be roughly classified into two categories and are both applied in medical domains for de-identification. The first uses grammar-based or rule-based techniques [1]. Such hand-crafted systems may require months of work by experienced domain experts and the rules will likely need to change for different data repositories. The second uses statistical learning approaches such as support vector machine (SVM)-based classification methods [16]. However, an SVM based method such as [16] only performs binary classification of the terms into protected health information (PHI) or non-PHI and does not allow statistical de-identification that requires the knowledge of different types of identifying attributes.

In our system, we use the statistical learning approach, in particular, a Conditional Random Fields-based named entity recognizer (NER), for extracting identifying and sensitive attributes. A conditional random field (CRF) [8] is an advanced discriminative probabilistic model that is shown to be effective in labeling natural language text. A CRF takes as input a sequence of tokens from the text where

[3]We focus on integrating structured and unstructured data. While semantic heterogeneity is an important challenge, it is not the focus of this paper.

each token has a feature set based on the sequence. Given a token from the sequence it calculates the probabilities of the various possible labels (whether it is a particular type of identifying or sensitive attribute) and chooses the one with maximum probability. The probability of each label is a function of the feature set associated with that token. More specifically, a CRF is an undirected graphical model that defines a single log-linear distribution function over label sequences given the observation sequence. The CRF is trained by maximizing the log-likelihood of the training data.

A key to the CRF classifier is the selection of the feature set. In our system, the features of a token contain previous word, next word, and properties such as capitalization, whether special characters exists, or if the token is a number, etc. The features we selected were largely influenced by suggestions in the recent executable survey of biomedical NER systems [9].

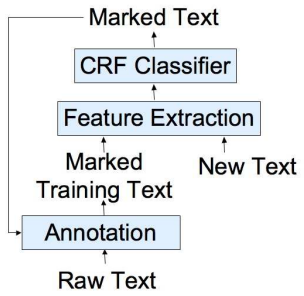

Figure 2: Iterative Annotation and Attribute Extraction Process

To facilitate the overall attribute extraction process, a unique feature of our approach is that it uses an iterative process for classifying and retagging, which allows the construction of a large training dataset without intensive human efforts. Figure 2 illustrates the iterative process. Concretely, our approach consists of: 1) a tagging interface which can be used to tag data with identifying and sensitive attributes to build the training dataset, 2) a CRF-based classifier[4] to classify terms from the text into multiple classes (different types of identifiers and sensitive attributes), and 3) a set of data preprocessing and post-processing strategies for extracting the features from text data for the classifier and feeding the classified data back to the tagging software for retagging and corrections.

2.3 Anonymization

Once the person-centric identifier view is generated after attribute extraction and linking, we provide a set of options for de-identifying the data.

Full De-identification. Information is considered fully de-identified by HIPAA if all of the identifiers (direct and indirect) have been removed and there is no reasonable basis to believe that the remaining information could be used to identify a person. The full de-identification option allows a user to remove all explicitly stated identifiers.

Partial De-identification. As an alternative to full de-identification, HIPAA makes provisions for a limited data set[5] from which direct identifiers (such as name and address) are removed, but not indirect ones (such as age). The partial de-identification option allows a user to remove the direct identifiers.

Statistical De-identification. Statistical de-identification attempts to maintain as much "useful" data as possible while guaranteeing statistically acceptable data privacy. Many such statistical criteria and anonymization techniques are proposed for structured data as we have discussed earlier. We adopt these techniques that allow a user to anonymize the data based on a chosen privacy principle.

Among the many privacy principles or criteria, k-anonymity [18] and its extension l-diversity [12] are the two most widely accepted and serve as the basis for many others. k-anonymity requires that a set of k records (entities) to be indistinguishable from each other based on a quasi-identifier set. An improved principle, l-diversity [12], demands every group to contain at least l well-represented sensitive values. Table I illustrates one possible anonymization of the original table with respect to the quasi-identifier set $(Age, Gender, Zipcode)$ that satisfies 2-anonymity and 2-diversity.

Table 1: Illustration of Anonymization

Name	Age	Gender	Zipcode	Diagnosis
Henry	25	Male	53710	Influenza
Irene	28	Female	53712	Lymphoma
Dan	28	Male	53711	Bronchitis
Erica	26	Female	53712	Influenza

Original Data

Name	Age	Gender	Zipcode	Diagnosis
*	[25 − 28]	Male	[53710-53711]	Influenza
*	[25 − 28]	Female	53712	Lymphoma
*	[25 − 28]	Male	[53710-53711]	Bronchitis
*	[25 − 28]	Female	53712	Influenza

Anonymized Data

More recently, t-closeness [11] is proposed to protect the numeric sensitive attributes that requires the distribution of sensitive values in each group to be analogous to the distribution of the entire dataset. δ-presence [15] is proposed to protect the presence of individuals in a published dataset. As these principles focus on "one-time" release of the data, m-invariance [20] is proposed to limit the risk of privacy disclosure in re-publication of the data.

Our current system includes the implementation of the Mondrian algorithm [10] that guarantees k-anonymity and an extended Incognito algorithm that guarantees l-diversity [12]. The Mondrian algorithm uses greedy recursive top-down partitioning of the (multidimensional) quasi-identifier domain space. It recursively chooses the split attribute with the largest normalized range of values, and (for continuous or ordinal attributes) partitions the data around the median value of the split attribute. This process is repeated until no allowable split remains, meaning that a particular region cannot be further divided without violating the anonymity constraint, or constraints imposed by value generalization hierarchies. The Incognito algorithm generates the set of all possible k-anonymous full-domain generalizations, with an optional tuple suppression threshold. Based on the subset property, the algorithm begins by checking single-attribute subsets of the quasi-identifier, and then iterates, checking k-anonymity and l-diversity with respect to increasingly large subsets. It is also on our research agenda to implement and incorporate more anonymization schemes that guarantees more advanced privacy principles.

[4]The Mallet toolkit [13] is used for the CRF implementation.

[5]limited data sets require data use agreements between the parties from which and to which information is provided.

3. DEMONSTRATION

In the demonstration, we will show the functionalities of our implemented system through real world data and highlight a few key features including: 1) the iterative process of data annotation and attribute extraction, and 2) the anonymization of the identifier view. We will also show some of the under-the-hood details for interested audience. The demonstration will be highly interactive such as allowing audience to mark a report, to add or remove certain features for attribute extraction, or to select different options for anonymization.

Dataset. To demonstrate the annotation and attribute extraction process we will use 100 textual pathology reports we collected in collaboration with Winship Cancer Institute at Emory University. For demonstration purposes, the real identifiers are replaced with artificial identifiers. In consultation with HIPAA compliance office at Emory, a small subset of the reports were tagged manually with identifiers including name, date of birth, age, zipcode, medical record numbers, and account numbers, and sensitive attributes such as diagnosis. To better demonstrate statistical de-identification that guarantees k-anonymity or l-diversity, we will also use artificial patient records with quasi-identifiers including age and zipcode. Figure 3 shows a sample pathology report section with personally identifying information such as age and medical record number highlighted.

CLINICAL HISTORY: 77 year old female with a history of B-cell lymphoma (Marginal zone, SH-02-22222, 6/22/01). Flow cytometry and molecular diagnostics drawn.

Figure 3: A Sample Pathology Report Section

Annotation and Attribute Extraction. We will show how to annotate reports using the iterative process with the feedback loop between manual and automatic annotation powered by a CRF-based classifier. We will show how the automatic annotation improves with feedback from the users and how the annotation becomes easier for the user as the CRF classifier is trained. Figure 4 shows a sample pathology report tagged with identifiers as the output of the attribute extraction component.

CLINICAL HISTORY: <Age>77</Age> year old <Gender>female</Gender> with a history of B-cell lymphoma (Marginal zone, <MRN>SH-02-22222</MRN>, 6/22/01). Flow cytometry and molecular diagnostics drawn.

Figure 4: A Sample Marked Report Section

De-identification. Once the identifying attributes are extracted and the reports are linked to each individual, the identifier view is generated from the original data. We show different de-identification options offered by the system. For full de-identification, we will remove all the identifying attributes. For partial de-identification, we only remove the direct identifiers including name and record numbers but do not remove indirect ones such as age. For statistical de-identification, we remove the direct identifiers and generalize attributes such as age and zipcode using the built-in anonymization algorithm. We will show before and after views of the de-identified reports. Figure 5 shows the sample de-identified pathology report as the output of the de-identification component.

*CLINICAL HISTORY: [70-79] year old female with a history of B-cell lymphoma (Marginal zone, **-**-****, 6/22/01). Flow cytometry and molecular diagnostics drawn.*

Figure 5: A Sample De-identified Report Section

Under-the-Hood. HIDE also provides a logging option during data operations. The demonstration interface will offer a view of the steps involved in de-identification so the audience will be able to see under-the-hood how the attribute extraction and anonymization proceed. Details include what features are generated from each report and how the attributes are selected for generalization.

4. REFERENCES

[1] R. M. B. A. Beckwith, U. J. Balis, and F. Kuo. Development and evaluation of an open source software tool for deidentification of pathology reports. *BMC Medical Informatics and Decision Making*, 6(12), 2006.

[2] R. J. Bayardo and R. Agrawal. Data privacy through optimal k-anonymization. In *ICDE*, 2005.

[3] E. Bertino, B. Ooi, Y. Yang, and R. H. Deng. Privacy and ownership preserving of outsourced medical data. In *ICDE*, 2005.

[4] B. C. M. Fung, K. Wang, and P. S. Yu. Top-down specialization for information and privacy preservation. In *ICDE*, 2005.

[5] J. Gardner and L. Xiong. HIDE: An integrated system for health information de-identification. In *CBMS*, 2008.

[6] D. Gupta, M. Saul, and J. Gilbertson. Evaluation of a deidentification (de-id) software engine to share pathology reports and clinical documents for research. *American Journal of Clinical Pathology*, 2004.

[7] P. Jurczyk, J. J. Lu, L. Xiong, J. D. Cragan, and A. Correa. FRIL: A tool for comparative record linkage. In *AMIA Annual Symposium*, 2008.

[8] J. Lafferty, A. McCallum, and F. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proc. 18th International Conf. on Machine Learning*, 2001.

[9] R. Leaman and G. G. Banner: An executable survey of advances in biomedical named entity recognition. In *Pacific Symposium on Biocomputing*, 2008.

[10] K. LeFevre, D. DeWitt, and R. Ramakrishnan. Mondrian multidimensional k-anonymity. In *IEEE ICDE*, 2006.

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

[12] A. Machanavajjhala, J. Gehrke, D. Kifer, and M. Venkitasubramaniam. l-diversity: Privacy beyond k-anonymity. In *ICDE*, 2006.

[13] A. K. McCallum. Mallet: A machine learning for language toolkit. http://mallet.cs.umass.edu, 2002.

[14] D. Nadeau and S. Sekine. A survey of named entity recognition and classification. *Linguisticae Investigationes*, 30(7), 2007.

[15] M. E. Nergiz, M. Atzori, and C. Clifton. Hiding the presence of individuals from shared databases. In *SIGMOD*, 2007.

[16] T. Sibanda and O. Uzuner. Role of local context in de-identification of ungrammatical fragmented text. In *North American Chapter of Association for Computational Linguistics/Human Language Technology*, 2006.

[17] L. Sweeney. Replacing personally-identifying information in medical records, the scrub system. *Journal of the American Informatics Association*, pages 333–337, 1996.

[18] L. Sweeney. k-anonymity: a model for protecting privacy. *International journal on uncertainty, fuzziness and knowledge-based systems*, 10(5), 2002.

[19] K. Wang, P. S. Yu, and S. Chakraborty. Bottom-up generalization: a data mining solution to privacy protection. In *ICDM*, 2004.

[20] X. Xiao and Y. Tao. M-invariance: towards privacy preserving re-publication of dynamic datasets. In *SIGMOD Conference*, pages 689–700, 2007.

[21] Q. Zhang, N. Koudas, D. Srivastava, and T. Yu. Aggregate query answering on anonymized tables. In *ICDE*, 2007.