



Contents lists available at ScienceDirect

## Data &amp; Knowledge Engineering

journal homepage: [www.elsevier.com/locate/datak](http://www.elsevier.com/locate/datak)

# An integrated framework for de-identifying unstructured medical data

James Gardner\*, Li Xiong

Department of Mathematics and Computer Science, Emory University

## ARTICLE INFO

## Article history:

Available online xxxxx

## Keywords:

Anonymization  
Medical text  
Named entity recognition  
Conditional random fields  
Cost-proportionate sampling  
Data linkage

## ABSTRACT

While there is an increasing need to share medical information for public health research, such data sharing must preserve patient privacy without disclosing any information that can be used to identify a patient. A considerable amount of research in data privacy community has been devoted to formalizing the notion of identifiability and developing techniques for anonymization but are focused exclusively on structured data. On the other hand, efforts on de-identifying medical text documents in medical informatics community rely on simple identifier removal or grouping techniques without taking advantage of the research developments in the data privacy community. This paper attempts to fill the above gaps and presents a framework and prototype system for de-identifying health information including both structured and unstructured data. We empirically study a simple Bayesian classifier, a Bayesian classifier with a sampling based technique, and a conditional random field based classifier for extracting identifying attributes from unstructured data. We deploy a  $k$ -anonymization based technique for de-identifying the extracted data to preserve maximum data utility. We present a set of preliminary evaluations showing the effectiveness of our approach.

© 2009 Elsevier B.V. All rights reserved.

## 1. Introduction

Current information technology enables many organizations to collect, store, and use various types of information about individuals. The government and organizations are increasingly recognizing the critical value of sharing such a wealth of information. However, individually identifiable information is protected under the Health Insurance Portability and Accountability Act (HIPAA).<sup>1</sup>

### 1.1. Motivating scenarios

The National Cancer Institute initiated the shared pathology informatics network (SPIN)<sup>2</sup> for researchers throughout the country to share pathology-based data sets annotated with clinical information to discover and validate new diagnostic tests and therapies. Fig. 1 shows a sample pathology report section with personally identifying information such as age and medical record number highlighted. It is necessary for each institution to de-identify or anonymize the data before having it accessible by the network. This network of shared data consists of both structured and unstructured data of various formats. Most medical data is heterogeneous meaning that even structured data from different institutions are labeled differently and

\* Corresponding author.

E-mail addresses: [jgardn3@emory.edu](mailto:jgardn3@emory.edu) (J. Gardner), [lxiong@emory.edu](mailto:lxiong@emory.edu) (L. Xiong).

<sup>1</sup> 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.

<sup>2</sup> Shared pathology informatics network. <http://www.cancerdiagnosis.nci.nih.gov/spin/>.

*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.*

**Fig. 1.** A sample pathology report section.

unstructured data is inherently heterogeneous. We use the terms heterogeneous and unstructured data interchangeably throughout this paper.

### 1.2. Existing and potential solutions

Currently, investigators or institutions wishing to use medical records for research purposes have three options: obtain permission from the patients, obtain a waiver of informed consent from their Institutional Review Boards (IRB), or use a data set that has had all or most of the identifiers removed. The last option can be generalized into the problem of de-identification or anonymization (both de-identification and anonymization are used interchangeably throughout this paper) where a *data custodian* distributes an anonymized view of the data that does not contain individually identifiable information to a (*data recipient*). It provides a scalable way for sharing medical information in large scale environments while preserving privacy of patients.

At the first glance, the general problem of data anonymization has been extensively studied in recent years in the data privacy community [11]. The seminal work by Sweeney et al. shows that a dataset that simply has identifiers removed is subject to linking attacks [35]. Since then, a large body of work contributes to data anonymization that transforms a dataset to meet a privacy principle such as *k*-anonymity using techniques such as generalization, suppression (removal), permutation and swapping of certain data values so that it does not contain individually identifiable information, such as [15,41,4,1,12,5,47,21,22,43,46].

While the research on data anonymization has made great progress, its practical utilization in medical fields lags behind. An overarching complexity of medical data, but often overlooked in data privacy research, is data heterogeneity. A considerable amount of medical data resides in unstructured text forms such as clinical notes, radiology and pathology reports, and discharge summaries. 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. Unfortunately, the bulk of data privacy research focus exclusively on structured data.

On the other hand, efforts on de-identifying medical text documents in medical informatics community [33,34,37,36,14,32,3,39] are mostly specialized for specific document types or a subset of HIPAA identifiers. Most importantly, they rely on simple identifier removal techniques without taking advantage of the research developments from data privacy community that guarantee a more formalized notion of privacy while maximizing data utility.

### 1.3. Contributions

Our work attempts to fill the above gaps and bridge the data privacy community and medical informatics community by developing a framework and prototype system, HIDE, for Health Information DE-identification of both structured and unstructured data. The contributions of our work are two fold. First, our system advances the medical informatics field by adopting information extraction (also referred to as attribute extraction) and data anonymization techniques for de-identifying heterogeneous health information. Second, the conceptual framework of our system advances the data privacy field by integrating the anonymization process for both structured and unstructured data. The specific components and contributions of our system are as follows:

- *Identifying and sensitive information extraction.* We leverage and empirically study existing named entity extraction techniques [25,30], in particular, simple Bayesian classifier and sampling based techniques, and conditional random fields based techniques to effectively extract identifying and sensitive information from unstructured data.
- *Data linking.* In order to preserve privacy for individuals and apply advanced anonymization techniques in the heterogeneous data space, we propose a structured *identifier view* with identifying attributes linked to each individual.
- *Anonymization.* We perform data suppression and generalization on the identifier view to anonymize the data with different options including full de-identification, partial de-identification, and statistical anonymization based on *k*-anonymization.

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 evaluate our prototype system through a set of real-world data and show the effectiveness of our approach.

In the rest of the paper we first describe related work. Then we describe our de-identification system including privacy models, the conceptual framework, identifier/attribute extraction, data linking, and anonymization. We then describe our experiments and results. Finally we conclude and describe further avenues of future work.

## 2. Related work

Our work is inspired and informed by a number of areas. We briefly review the most relevant areas below and discuss how our work leverages and advances the current state-of-the-art.

### 2.1. Privacy preserving data publishing

Privacy preserving data publishing for centralized databases has been studied extensively in recent years. One thread of work aims at devising privacy principles, such as  $k$ -anonymity and later principles that remedy its problems, that serve as criteria for judging whether a published dataset provides sufficient privacy protection [35,24,38,2,23,44,26,31,8,7]. Another large body of work contributes to algorithms that transforms a dataset to meet one of the above privacy principles (dominantly  $k$ -anonymity) [15,41,29,4,1,12,5,21,22,40,18,43,46,8,7]. The bulk of this work has focused exclusively on structured data.

### 2.2. Medical text de-identification

In the medical informatics community, there are some efforts on de-identifying medical text documents [33,34,37,36,14,32,3,39]. Most of them uses a two-step approach which extracts the identifying attributes first and then removes or masks the attributes for de-identification purposes. Most of them are specialized for specific document types (e.g. pathology reports only [37,14,3]). Some of them focus on a subset of HIPAA identifiers (e.g. name only [36,37]) while some others focus on differentiating protected health information (PHI) from non-PHI [32]. Most importantly, most of these work rely on simple identifier removal or grouping techniques and do not take advantage of the recent research developments that guarantee a more formalized notion of privacy while maximizing data utility.

### 2.3. Information extraction

Extracting atomic identifying and sensitive attributes (such as name, address, and disease name) from unstructured text such as pathology reports can be seen as an application of named entity recognition (NER) problem [25,30]. NER systems can be roughly classified into two categories and both are applied in medical domains for de-identification. The first uses grammar-based or rule-based techniques [3]. Unfortunately such hand-crafted systems may take the cost of 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. However, an SVM based method such as [32] only performs binary classification of the terms into PHI or non-PHI and does not allow statistical de-identification that requires the knowledge of different types of identifying attributes.

## 3. De-identification system

We first present the privacy and de-identification models used in our system, then present the conceptual framework behind our system, followed by a discussion on each component with its research challenges and proposed solutions.

### 3.1. Privacy model

Protected health information (PHI) is defined by HIPAA as individually identifiable health information. Identifiable information refers to data explicitly linked to a particular individual as well as data that could enable individual identification. Personal identifiers include direct ones such as name and Social Security number as well as indirect ones such as age, gender, address information, etc. We adopt the following privacy models or de-identification options in our framework.

#### 3.1.1. 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. While the explicitly stated identifiers can be removed, the final category of HIPAA identifiers includes “any other unique identifying number, characteristic, or code” and makes it nearly impossible to guarantee with absolute certainty that data is fully de-identified. In addition, a full de-identification would render the data not very useful for many data analysis purposes.

#### 3.1.2. Partial de-identification

As an alternative to full de-identification, HIPAA makes provisions for a limited data set<sup>3</sup> from which direct identifiers (such as name and address) are removed, but not indirect ones (such as age). This approach provides better data utility.

<sup>3</sup> Limited data sets require data use agreements between the parties from which and to which information is provided.

### 3.1.3. 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 de-identification techniques are proposed for structured data as we have discussed earlier. Our approach generalizes these notions to heterogeneous data and we will discuss them in detail as we discuss the de-identification techniques in a later subsection.

### 3.2. Conceptual framework

The general conceptual framework of our system consists of a number of key components that integrate de-identification for a heterogeneous data space utilizing advanced anonymization schemes. Fig. 2 presents an illustration of the framework. We present an overview below and give more details on the important components in subsequent subsections.

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 HIPAA identifiers as well as sensitive attributes from unstructured data. Note that in order to apply advanced data anonymization techniques, this will be a much broader set of information to be extracted than existing de-identification systems that typically focus on the set or a subset of HIPAA identifiers.

In relational data, we assume each tuple corresponds to an individual entity. This mapping is not present in heterogeneous medical data repositories. For example, one patient may have multiple pathology and lab reports prepared at different times. In order to preserve privacy for individuals and apply static de-identification in this complex data space, the *data linking* component links relevant attributes (structured or extracted) to each individual entity and produces a person-centric representation of the data.

Once the identifying attributes are extracted and linked to individuals, they form a structured *identifier view*. This notion of identifier view will allow application of advanced anonymization algorithms that are otherwise not applicable to unstructured data. Given an identifier view, 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.

### 3.3. Attribute extraction

Our first challenge is to recognize and extract identifying as well as sensitive information from the unstructured data in order to apply advanced anonymization algorithms on the heterogeneous data. While information extraction is a challenging research field of its own, we leverage the current research results from the information extraction field, present the approach we take, as well as specific research challenges within our context.

We use a statistical learning approach for extracting identifying and sensitive attributes. Note that we aim at a much broader set of attributes than existing de-identification systems which only focus on the set of a subset of the HIPAA identifiers. To facilitate the overall attribute extraction process, a unique aspect of our approach is that it uses an iterative process

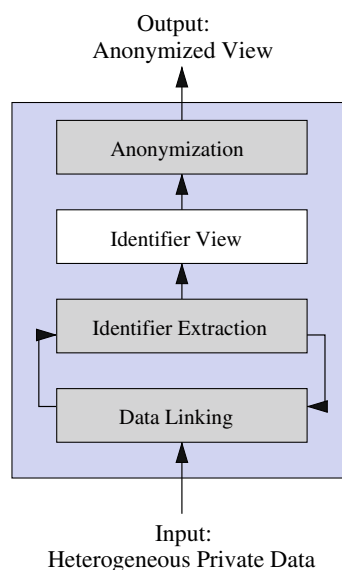


Fig. 2. Conceptual framework.

for classifying and retagging which allows the construction of a large training dataset without intensive human efforts in labeling the data from scratch. Fig. 3 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 feature generation component for extracting the features from text data for the classifier, (3) a classifier to classify terms from the text into multiple classes (different types of identifiers and sensitive attributes), and (4) a set of data postprocessing strategies for feeding the classified data back to the tagging software for retagging and corrections.

A key to our classifier based approach is the selection of the feature set. Once the text is parsed, a set of features is generated for each token or term in the text. In our current system, the features of a token contain the token itself, previous word, next word, and things such as capitalization, whether special characters exists, or if the token is a number, etc. The features we used were largely influenced by suggestions in the recent executable survey of biomedical NER systems [20]. It is possible to include the use of medical ontologies for extracting attributes but it seems that our system achieves good results by using local features (features about the words and surrounding words) without having to result to using global features (features about or relative to the entire dataset). The use of local features allow our system to be more portable and work across many different types of data.

Once the feature data is generated, we feed them to a classifier for training. The learned classifier is then used to tag new text. We discuss below two classification approaches we studied.

### 3.3.1. Conditional random fields based classification

We first adopted a conditional random fields based named entity recognizer (NER) for extracting identifying and sensitive attributes. A conditional random field (CRF) [19] 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 each token has a feature set based on the sequence. Given a token from the sequence it calculates the probabilities of the various possible labeling (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. The Mallet toolkit [28] is used for the CRF implementation.

### 3.3.2. Prioritized classification with cost-proportionate sampling

One of the drawback of the CRF classifier is its long training time. We also experimented with a simple Naive Bayesian classifier on the feature set we have generated. As expected (results will be presented and discussed in next Section), the Bayesian classifier performs poorly, specially for the relatively rare types of identifying and sensitive attributes. This is mainly due to the fact that the non-identifying terms (or the terms with *other* class label in our classification system) comprise more than 99% of the total terms and hence the prior probability for most of the identifying attributes are extremely small. In addition, the classifier missed quite a large portion of directly identifying attributes such as names. This is considered detrimental compared to a classifier that misses a few indirect identifiers such as age or address attributes. In general, different cost (risk of identifying a person) can be associated with failing to de-identify certain individual types attributes.

The above observations motivate us to develop a *prioritized* classification approach through a cost-proportionate sampling technique [45]. The basic idea is that random examples from the original dataset (the feature set of all tokens in our case) are chosen and added to the training set based on specified probability for each instance. The probability of being added to the training set is based on the class label of the instance. By assigning different probabilities to different class la-

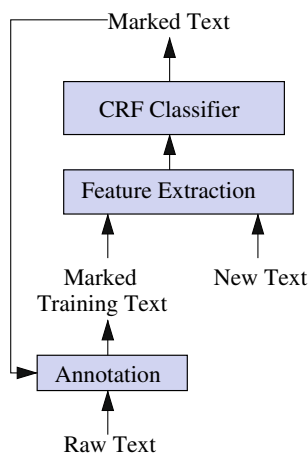


Fig. 3. Attribute extraction process.

bels we force our training set to contain more or less instances of particular classes. For example, since missing a name attribute incurs a higher cost or a higher risk of identification of individuals, we assign a higher probability for the name class. As a result, tokens that are tagged as a name, have a higher probability making it into the training set, and hence boost the attribute extraction accuracy for the name attribute.

### 3.4. Data linking

In order to build an identifier view of the data, we need to link all relevant information (including identifiers, quasi-identifiers, and sensitive information) to each individual entity. The task is simple when there is a unique identifier that links all the relevant documents and structured information about a person in the data repository. However, this unique identifier may not be present in many cases. This problem is quite reminiscent of and in fact relevant to the traditional record linkage problem. However, the heterogeneous data also presents a number of new challenges to the linking problem in our context. While the traditional record linkage problem assumes structured data where one reference (record) refers to an entity and there are several attributes associated with every reference, the data can be heterogeneous in our context. An entity may contain any number of attributes that can be extracted from the unstructured data. In addition, the associations between attributes and references (entities) are not clearly defined in the heterogeneous data.

Our approach includes an iterative two-step solution involving data linking as well as attribute extraction. The extraction component extracts relevant attributes from the text and link or add them to the existing or new entities in our database. The linking component links or merges the tuples based on the structured and extracted attributes using existing record linkage techniques. For instance, two tuples with the same or similar name and demographic information but with different lab reports generated at different times will be merged. It is important to note that the two steps can be performed in a *repeated and iterative* manner so the data linking and attribute extraction are refined and enriched over iterations. For example, the linked information can be used in the extraction process to improve the extraction accuracy and extract new additional attributes.

For linking structured and extracted attributes, probabilistic record linkage techniques [13,42] are used to resolve the potential attribute conflicts and possible errors introduced in the attribute extraction process. In particular the three step approach is being used (1) a vector of similarity scores is computed for individual record pairs by comparing their corresponding attributes, (2) each reference pair is classified as either a match or non-match based on the similarity score, and (3) a transitive closure is computed over matching pairs. We adopted a fine-grained record integration and linkage tool we developed, FRIL [16], for this purpose. We refer readers to [16] for the algorithmic and implementation details of the system.

We note that our extraction component so far only extracts atomic attributes and we use simple heuristics to associate these attributes to an entity. For example, it is likely that all identifying attributes extracted from the same pathology report such as age and address correspond to the same patient. But in many cases, the simple heuristics could fail and the problem is much more challenging. We will explore value correlation techniques such as [9] as well as the idea of dependencies or associations between linkage decisions [6,17,10] in our future research agenda.

### 3.5. Anonymization

Once the identifier view is generated after attribute extraction and linking, we can perform attribute removal (suppression) to allow full de-identification (as possible) and partial de-identification. We also allow statistical de-identification through anonymization techniques through attribute generalization that guarantees privacy based on a privacy principle while maintaining maximum data utility. Among the many privacy principles or criteria, *k*-anonymity [35] and its extension *l*-diversity [24] are the two most widely accepted and serve as the basis for many others, and hence, are used in our initial work. Below we illustrate the basic ideas behind these principles and present the anonymization approach we used.

In defining anonymization given a relational table *T*, the attributes are characterized into three types. *Unique identifiers* are attributes that identify individuals. *Quasi-identifier set* is a minimal set of attributes that can be joined with external information to re-identify individual records. We assume that a quasi-identifier is recognized based on the domain knowledge. *Sensitive attributes* are those attributes that an adversary should not be permitted to uniquely associate their values with a unique identifier. Table 1 illustrates an original relational table of personal information where *Name* is considered as an identifier, (*Age*, *Gender*, *Zipcode*) a quasi-identifier set, and *Diagnosis* a sensitive attribute.

The *k*-anonymity model provides an intuitive requirement for privacy in that no individual record should be uniquely identifiable from a group of *k* with respect to the quasi-identifier set. The set of all tuples in *T* containing identical values for the quasi-identifier set is referred to as *equivalence class*. *T* is *k*-anonymous if every tuple is in an equivalence class of size at least *k*. A *k*-anonymization of *T* is a transformation or generalization of the data *T* such that the transformed dataset is *k*-anonymous. The *l*-diversity model provides an extension to *k*-anonymity and requires that each equivalence class also contains at least *l* well-represented distinct values for a sensitive attribute to avoid the homogeneous sensitive information revealed for the group. Table 1 illustrates one possible anonymization with respect to the quasi-identifier set (*Age*, *Gender*, *Zipcode*) that satisfies 2-anonymity and 2-diversity.

A large number of algorithms have been developed for structured data anonymization based on a certain privacy principle (dominantly *k*-anonymity). In this study, we adopt the Mondrian multidimensional approach [22] which is a *k*-anonymiza-

**Table 1**  
Illustration of anonymization.

Name	Age	Gender	Zipcode	Diagnosis
<i>Original data</i>				
Henry	25	Male	53710	Influenza
Irene	28	Female	53712	Lymphoma
Dan	28	Male	53711	Bronchitis
Erica	26	Female	53712	Influenza
<i>Anonymized data</i>				
*	[25–28]	Male	[53710–53711]	Influenza
*	[25–28]	Female	53712	Lymphoma
*	[25–28]	Male	[53710–53711]	Bronchitis
*	[25–28]	Female	53712	Influenza

tion algorithm that has been shown to have advantages compared to others. The Mondrian algorithm uses greedy recursive top-down partitioning of the (multidimensional) quasi-identifier domain space. In order to obtain approximately uniform partition occupancy, it recursively chooses the split attribute with the largest normalized range of values, referred to as *spread*, 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.

#### 4. Experiments

We conducted a set of preliminary experiments on a real-world dataset. In this section, we first describe our dataset and experiment setup and then present the preliminary results demonstrating the effectiveness of our approach.

##### 4.1. Dataset and experiment setup

Our dataset contains 100 textual pathology reports we collected in collaboration with Winship Cancer Institute at Emory. In consultation with HIPAA compliance office at Emory, the reports were tagged manually with identifiers including name, date of birth, age, medical record numbers, and account numbers or *other* if the token was not one of the identifying attributes. There were in total 106,255 token-tag pairs in our dataset. The tagging process involved initial tagging of a small set of reports, automatic tagging for the rest of the reports with our attribute extraction component using the small training set, and manual retagging or correction for all the reports.

We used the dataset for evaluating the accuracy of our attribute extraction component (discussed in Section 3.3). Fig. 4 shows a sample pathology report tagged with identifiers as the output of the attribute extraction component.

We compared the Naive Bayes on the original dataset, Naive Bayes with cost-proportionate rejection sampling, and the CRF approach. For cost-proportionate sampling, Table 2 shows the probabilities we used with each type of attributes. We generated a file with 200,000 examples using the sampling from the original feature file with 106,255 examples.

Once the identifying attributes are extracted and the reports are linked to each individual, we applied different de-identification options on the original dataset. For full de-identification, we removed all the identifying attributes.

For partial de-identification, we only removed the direct identifiers including name and record numbers but did not remove indirect ones such as age. For statistical de-identification, we removed the direct identifiers and generalized age attribute using the *k*-anonymization algorithm built in our anonymization component (discussed in Section 3.5). Fig. 5 shows the sample de-identified pathology report as the output of the statistical de-identification component.

We then evaluated the utility of the anonymized data through a set of queries.

##### 4.2. Attribute extraction

To evaluate the effectiveness of our attribute extraction component, we conducted a set of experiments using ten fold cross-validation in which the dataset was divided into 10 subsets and 9 subsets were used for training and the other 1 was used for testing and it was repeated 10 times (once for each subset).

*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.*

**Fig. 4.** A sample marked report section.

**Table 2**

Probability values used in cost-proportionate sampling.

Label	Probability
Medical record number	.2
Account number	.2
Age	.3
Date	.5
Name (begin)	1
Name (intermediate)	1
Other	.1

*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.*

**Fig. 5.** A sample de-identified report section.

#### 4.2.1. Metrics

We report *precision*, *recall* as well as the *F1 metric* for our experiments. Precision ( $P$ ) or the positive predictive value is defined as the number of correctly labeled identifying attributes over the total number of labeled identifying attributes. Recall ( $R$ ) is defined as the number of correctly labeled identifying attributes over the total number of identifying attributes in the text, and  $F1$  is defined as  $F1 = 2(P \cdot R)/(P + R)$ . It is worth noting that *sensitivity* is defined the same as recall and *specificity* is defined as the number of correctly labeled non-identifying attributes over the total number of non-identifying attributes in the text. We do not report specificity because the non-identifying attributes are dominating compared to the identifying attributes so specificity will be always close to 100% which would not be very informative.

#### 4.2.2. Results

Tables 3–5 presents the extraction results in precision, recall and  $F1$  metric for each identifying attribute (class) as well as the overall accuracy for each of the extraction techniques, respectively. We observe that the results from the Naive Bayes with biased rejection sampling are much better than those without the biased rejection sampling. The results for Naive Bayes with biased rejection sampling are comparable or even better than the CRF-based classifier for certain attributes. This is somewhat surprising to us considering the simplicity of Bayesian and complexity of the CRF classifier. We suspect that the good result achieved by the Bayesian method is largely due to the sampling technique and the fairly homogeneous dataset we have but the result is yet to be generalized and confirmed with other datasets.

In general, the CRF approach achieves the best overall result. In particular, it is much better at detecting account number, which neither Naive Bayes approach ever detects. While finding proper names (and how long those names extend) can be still further improved, most attributes achieve nearly perfect performance. The effectiveness is largely contributed to the well developed CRF method and the relevant features shown useful for personal health information (PHI) extraction as well as the relatively homogeneous data format in our dataset. We plan to add new features, feature induction [27], part of speech tagging to further improve the performance for various datasets.

#### 4.3. De-identification

In many public health and outcome research studies, a key step involves sub-population identification where researchers may wish to study a certain demographic population, such as males over 50, and learn classification models based on demographic information and clinical symptoms to predict diagnosis. To evaluate the effectiveness of different de-identification options, we ran a set of queries for a sub-population selection on the de-identified dataset and measured the query precision

**Table 3**

Attribute extraction accuracy using Naive Bayes.

Label	Precision	Recall	$F1$
<i>Overall accuracy: 0.75</i>			
Medical record number	0.915	0.9627	0.938
Account number	0	0	0
Age	1	0.5223	0.6802
Date	1	1	1
Name (begin)	1	0.9746	0.987
Name (intermediate)	1	0.4053	0.5754

**Table 4**

Attribute extraction accuracy using Naive Bayes with prioritized sampling.

Label	Precision	Recall	F1
<i>Overall accuracy: 0.98</i>			
Medical record number	0.9176	0.9962	0.9552
Account number	0	0	0
Age	1	0.9924	0.9963
Date	1	1	1
Name (begin)	1	1	1
Name (intermediate)	1	1	1

**Table 5**

Attribute extraction accuracy using CRF.

Label	Precision	Recall	F1
<i>Overall accuracy: 0.98</i>			
Medical record number	1.000	0.988	0.994
Account number	0.990	1.000	0.995
Age	1.000	0.963	0.981
Date	1.000	1.000	1.000
Name (begin)	0.970	0.970	0.970
Name (intermediate)	1.000	0.980	0.990

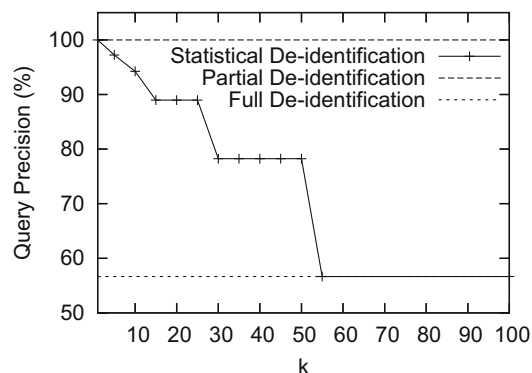
defined as % of correct reports being returned. Concretely, we randomly generated 10,000 queries with a selection predicate of the form  $age > n$  and  $age < n$  to select the corresponding reports (patients). Given a selection predicate  $age > 45$ , a report with age attribute anonymized to the range [40–50] would also be returned. Thus the query result gives perfect recall but varying precision and we report the query precision below.

Fig. 6 presents the query precision on the de-identified dataset using different de-identification options with varying  $k$  in  $k$ -anonymization based statistical de-identification. It can be observed that partial de-identification offers 100% precision as it did not de-identify age attribute. However, such de-identification provides limited data protection. On the other hand, full de-identification provides the maximum privacy protection, but suffers a low query precision. Statistical de-identification offers a tradeoff that provides a guaranteed privacy level while maximizing the data utility. As expected, the larger the  $k$ , the better the privacy level and the lower the query precision as the original data are generalized to a larger extent.

## 5. Conclusion and future works

We presented a conceptual framework as well as a prototype system for anonymizing heterogeneous health information including both structured and unstructured data. Our initial experimental results show that our system effectively detects a variety of identifying attributes with high precision, and provides flexible de-identification options that anonymizes the data with a given privacy guarantee while maximizing data utility to the researchers. While our work is a convincing proof-of-concept, there are several aspects that will be further explored.

First, we are exploring innovative anonymization approaches that prioritize the attributes based on how important and critical they are to the privacy preserving requirements as well as the application needs. Second, in addition to enhance the

**Fig. 6.** Query precision using different de-identification options.

(atomic) attribute extraction accuracy, a more in-depth and challenging problem that we will investigate is to extract indirect identifying information. For example, progeria is a very rare condition associated with unnaturally fast aging and simply knowing that a report concerns a patient with this condition makes an identification likely even if other identifiers are removed. Finally, we are planning to deploy the developed framework in the cancer patient data warehouse. Integration of the developed techniques into the Cancer Biomedical Informatics Grid (caBIG)<sup>4</sup> will also be carried out.

## Acknowledgements

This research is partially supported by an Emory URC grant and Emory ITSC grant. We thank the guest editors and anonymous reviewers for their valuable comments that improved this paper.

## References

- [1] C.C. Aggarwal, On  $k$ -anonymity and the curse of dimensionality, in: Thirty-first International Conference on Very Large Databases (VLDB), 2005, pp. 901–909.
- [2] G. Aggarwal, T. Feder, K. Kenthapadi, S. Kuller, R. Panigrahy, D. Thomas, A. Zhu, Achieving anonymity via clustering, in: Proceedings of the Twenty-Fifth ACM SIGACT–SIGMOD–SIGART Symposium on Principles of Database Systems, 2006, pp. 153–162.
- [3] R.M.B.A. Beckwith, U.J. Balis, 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).
- [4] R.J. Bayardo, R. Agrawal, Data privacy through optimal  $k$ -anonymization, in: ICDE'05: Proceedings of the 21st International Conference on Data Engineering (ICDE'05), Washington, DC, USA, IEEE Computer Society, 2005, pp. 217–228.
- [5] E. Bertino, B. Ooi, Y. Yang, R.H. Deng, Privacy and ownership preserving of outsourced medical data, in: Proceedings of the 21st International Conference on Data Engineering, ICDE 2005, 2005.
- [6] I. Bhattacharya, L. Getoor, Iterative record linkage for cleaning and integration, in: DMKD'04: Proceedings of the 9th ACM SIGMOD Workshop on Research issues in Data Mining and Knowledge Discovery, 2004.
- [7] Y. Bu, A. Fu, R. Wong, L. Chen, J. Li, Privacy preserving serial data publishing by role composition, in: Thirty-fourth International Conference on Very Large Data Bases (VLDB), 2008.
- [8] G. Cormode, D. Srivastava, T. Yu, Q. Zhang, Anonymizing bipartite graph data using safe groupings, in: Thirty-fourth International Conference on Very Large Data Bases (VLDB), 2008.
- [9] A. Culotta, A. McCallum, J. Betz, Integrating probabilistic extraction models and data mining to discover relations and patterns in text, in: HLT/NAACL, Morristown, NJ, USA, Association for Computational Linguistics, 2006, pp. 296–303.
- [10] X. Dong, A. Halevy, J. Madhavan, Reference reconciliation in complex information spaces, in: SIGMOD'05: Proceedings of the 2005 ACM SIGMOD International Conference on Management of Data, 2005.
- [11] B.C.M. Fung, K. Wang, R. Chen, P.S. Yu, Privacy-preserving data publishing: a survey on recent developments, *ACM Computing Surveys*, 2010.
- [12] B.C.M. Fung, K. Wang, P.S. Yu, Top-down specialization for information and privacy preservation, in: Proceedings of the 21st IEEE International Conference on Data Engineering (ICDE 2005), Tokyo, Japan, 2005, pp. 205–216.
- [13] L. Gu, R. Baxter, D. Vickers, C. Rainsford, Record linkage: current practice and future directions.
- [14] D. Gupta, M. Saul, 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) 76–186.
- [15] V.S. Iyengar, Transforming data to satisfy privacy constraints, in: Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2002, pp. 279–288.
- [16] P. Jurczyk, J.J. Lu, L. Xiong, J.D. Cragan, A. Correa, Fril: a tool for comparative record linkage, in: AMIA 2008 Annual Symposium, 2008.
- [17] D.V. Kalashnikov, S. Mehrotra, Z. Chen, Exploiting relationships for domain-independent data cleaning, in: SIAM International Conference on Data Mining, 2005.
- [18] D. Kifer, J. Gehrke, Injecting utility into anonymized datasets, in: SIGMOD Conference, 2006, pp. 217–228.
- [19] J. Lafferty, A. McCallum, F. Pereira, Conditional random fields: probabilistic models for segmenting and labeling sequence data, in: Proceedings of the 18th International Conference on Machine Learning, 2001.
- [20] R. Leaman, G.G. Banner, An executable survey of advances in biomedical named entity recognition, in: Pacific Symposium on Biocomputing, 2008.
- [21] K. LeFevre, D. Dewitt, R. Ramakrishnan, Incognito: efficient full-domain  $k$ -anonymity, in: ACM SIGMOD International Conference on Management of Data, 2005.
- [22] K. LeFevre, D. DeWitt, R. Ramakrishnan, Mondrian multidimensional  $k$ -anonymity, in: Proceedings of the 22nd International Conference on Data Engineering, ICDE 2006, 2006.
- [23] N. Li, T. Li,  $T$ -closeness: privacy beyond  $k$ -anonymity and  $l$ -diversity, in: Proceedings of the 23rd International Conference on Data Engineering, ICDE 2007, 2007.
- [24] A. Machanavajjhala, J. Gehrke, D. Kifer, M. Venkatasubramanian,  $L$ -diversity: privacy beyond  $k$ -anonymity, in: Proceedings of the 22nd International Conference on Data Engineering (ICDE'06), 2006, pp. 24.
- [25] C. Manning, H. Schütze, Foundations of Statistical Natural Language Processing, MIT Press, 1999.
- [26] D.J. Martin, D. Kifer, A. Machanavajjhala, J. Gehrke, J.Y. Halpern, Worst-case background knowledge for privacy-preserving data publishing, in: Proceedings of the 23rd International Conference on Data Engineering, ICDE 2007, 2007.
- [27] A. McCallum, Efficiently inducing features of conditional random fields, in: 19th Conference in Uncertainty in Artificial Intelligence (UAI), 2003.
- [28] A.K. McCallum, Mallet: a machine learning for language toolkit. <<http://mallet.cs.umass.edu>>, 2002.
- [29] A. Meyerson, R. Williams, On the complexity of optimal  $k$ -anonymity, in: Proceedings of the Twenty-third ACM SIGACT–SIGMOD–SIGART Symposium on Principles of Database Systems, 2004, pp. 223–228.
- [30] D. Nadeau, S. Sekine, A survey of named entity recognition and classification, *Linguisticae Investigationes* 30 (7) (2007).
- [31] M.E. Nergiz, M. Atzori, C. Clifton, Hiding the presence of individuals from shared databases, in: SIGMOD Conference, 2007, pp. 665–676.
- [32] T. Sibanda, 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.
- [33] L. Sweeney, Replacing personally-identifying information in medical the records scrub system, *Journal of the American Informatics Association* (1996) 333–337.
- [34] L. Sweeney, Guaranteeing anonymity when sharing medical data, the datafly system, in: Proceedings of AMIA Annual Fall Symposium, 1997.
- [35] L. Sweeney,  $k$ -Anonymity: a model for protecting privacy, *International Journal on Uncertainty Fuzziness, and Knowledge-based Systems* 10 (5) (2002).
- [36] R.K. Taira, A.A. Bui, H. Kangaroo, Identification of patient name references within medical documents using semantic selectional restrictions, in: Proceedings of AMIA Symposium, 2002, pp. 757–761.

<sup>4</sup> Cancer Biomedical Informatics Grid. <https://cabig.nci.nih.gov/>.

- [37] S.M. Thomas, B. Mamlin, G.S. Adn, C. McDonald, A successful technique for removing names in pathology reports, in: Proceedings of AMIA Symposium, 2002, pp. 777–781.
- [38] T.M. Truta, B. Vinay, Privacy protection:  $p$ -sensitive  $k$ -anonymity property, in: Proceedings of the 22nd International Conference on Data Engineering Workshops, ICDE 2006, 2006, pp. 94.
- [39] O. Uzuner, Y. Luo, P. Szolovits, Evaluating the state-of-the-art in automatic de-identification, Journal of the American Medical Informatics Association 14 (5) (2007).
- [40] K. Wang, B.C.M. Fung, Anonymizing sequential releases, in: Proceedings of the Twelfth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2006.
- [41] K. Wang, P.S. Yu, S. Chakraborty, Bottom-up generalization: a data mining solution to privacy protection, in: Proceedings of the 4th IEEE International Conference on Data Mining (ICDM 2004), November 2004.
- [42] W. Winkler, Overview of record linkage and current research directions, Technical Report Statistics #2006-2, Statistical Research Division, US Bureau of the Census, 2006.
- [43] X. Xiao, Y. Tao, Anatomy: simple and effective privacy preservation, in: Thirty-second International Conference on Very Large Databases (VLDB), 2006, pp. 139–150.
- [44] X. Xiao, Y. Tao,  $M$ -invariance: towards privacy preserving re-publication of dynamic datasets, in: SIGMOD Conference, 2007, pp. 689–700.
- [45] B. Zadrozny, J. Langford, N. Abe, Cost-sensitive learning by cost-proportionate example weighting, in: ICDM'03: Proceedings of the Third IEEE International Conference on Data Mining, Washington, DC, USA, IEEE Computer Society, 2003.
- [46] Q. Zhang, N. Koudas, D. Srivastava, T. Yu, Aggregate query answering on anonymized tables, in: Proceedings of the 23rd International Conference on Data Engineering, ICDE 2007, 2007, pp. 116–125.
- [47] S. Zhong, Z. Yang, R.N. Wright, Privacy-enhancing  $k$ -anonymization of customer data, in: Proceedings of the Twenty-fourth ACM SIGACT–SIGMOD–SIGART Symposium on Principles of Database Systems, 2005.



**James Gardner** received his BS in Mathematics and Computer Science from East Tennessee State University and his MS in Computer Science from Emory University. He is currently pursuing his Ph.D. at Emory. His current research is focused on Machine Learning, Natural Language Processing, and Semantic Web Technologies.



**Li Xiong** is an Assistant Professor of Mathematics and Computer Science at Emory University. She holds a Ph.D. from Georgia Institute of Technology and an MS from Johns Hopkins, both in Computer Science. She also worked as a software engineer in IT industry for several years prior to pursuing her doctorate. Her areas of interests are in data and information management, data privacy and security, and bio and health informatics. She is a recipient of a Career Enhancement Fellowship from the Woodrow Wilson Foundation.