HIDE: An Integrated System for Health Information DE-identification*

James Gardner and Li Xiong
Department of Mathematics and Computer Science
Emory University
{jgardn3, lxiong}@emory.edu

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 identifiable information. 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 prototype system for de-identifying health information including both structured and unstructured data. It deploys a conditional random fields based technique for extracting identifying attributes from unstructured data and k-anonymization based technique for de-identifying the data while preserving maximum data utility. We present a set of preliminary evaluations showing the effectiveness of our approach.

1. Introduction

Current information technology enables many organizations to collect, store, and use various types of medical information about individuals. Government and organizations increasingly recognize the critical value in sharing such a wealth of information. However, individually identifiable information is protected under the Health Insurance Portability and Accountability Act (HIPAA)[1] and other laws and policies.

Motivating Scenarios. The National Cancer Institute initiated the Shared Pathology Informatics Network (SPIN)[2] for researchers throughout the country to share pathology-based data sets annotated with clinical information to discover and validate new diagnostic tests and therapies. Figure 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.

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 1. A Sample Pathology Report Section

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 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 data privacy community. The seminal work by Sweeney et al. shows that a dataset that simply has identifiers removed is subject to linking attacks [27]. A few *principles* including k-anonymity and later principles that remedy its problems have been proposed that serve as criteria for judging whether a published dataset provides sufficient privacy protection [27, 16, 30, 2, 15, 34, 18, 23]. 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 [8, 32, 21, 4, 1, 6, 5, 36, 12, 13, 14, 31, 9, 33, 35].

*This research is partially supported by an Emory URC grant and an Emory ITSC grant.

[1] Health Insurance Portability and Accountability Act (HIPAA). http://www.hhs.gov/ocr/hipaa/.

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

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.

In the medical informatics community, there are some efforts on de-identifying medical text documents [25, 26, 29, 28, 7, 24, 3]. Most of them are specialized for specific document types (e.g. pathology reports only [29, 7, 3]). Some of them focus on a subset of HIPAA identifiers (e.g. name only [28, 29]) while some others focus on differentiating Protected Health Information (PHI) from non-PHI [24]. 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.

Contributions. Our work attempts to fill the above gaps and bridge the data privacy community and medical informatics community by developing a 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 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: 1) *Data Linking*. In order to preserve privacy for individuals and apply advanced anonymization techniques in the heterogeneous data space, we propose a structured *person-centric identifier view* with identifying attributes linked to each individual, 2) *Identifying and Sensitive Information Extraction*. We leverage the latest Named Entity Extraction techniques [17, 22], in particular, conditional random fields based techniques to effectively extract identifying and sensitive information from unstructured data, and 3) *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. Finally, we evaluate our prototype system through a set of real-world data and show the effectiveness of our approach.

2 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.

2.1 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.

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.

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

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.

2.2 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. Figure 2 presents an illustration of the framework. We present an overview below

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

and give more details on the important components in subsequent subsections.

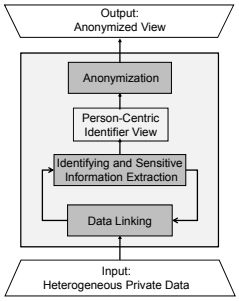

Figure 2. Conceptual Framework

In relational data, we assume each tuple corresponds to an individual entity. This mapping is not present in heterogeneous medical data repository. For example, one patient may have multiple pathology and lab reports prepared at different times. In order to preserve privacy for individuals and apply statical 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.

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.

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.

2.3 Attribute 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 [17, 22]. 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 [24] 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.

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) [10] 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.

A key to the CRF classifier is the selection of the feature set. In our system, the features of a token contains previous word, next word, and things 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 [11].

To facilitate the overall attribute extraction process, our approach consists of: 1) a tagging software which can be used to tag data with identifying and sensitive attributes to build the training dataset, 2) a CRF-based classifier to classify terms from the text into multiple classes (different types of identifiers and sensitive attributes), and 3) a set of data preprocessing and postprocessing 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. A unique feature of our approach is its iterative process for classifying and retagging which allows the construction of a large training dataset without intensive human efforts in labeling the data from scratch. The Callisto annotation tool[4] is used for the tagging and annotation implementation. The Mallet toolkit [20] is used for the CRF implementation.

[4]Callisto annotation tool. http://callisto.mitre.org/

2.4 Anonymization

Once the person-centric 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 [27] and its extension l-diversity [16] 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.

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	Disease
$*$	$[25-28]$	Male	[53710-53711]	Influenza
$*$	$[25-28]$	Female	53712	Lymphoma
$*$	$[25-28]$	Male	[53710-53711]	Bronchitis
$*$	$[25-28]$	Female	53712	Influenza

Anonymized Data

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-identifer 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 [13] which is a k-anonymization algorithm that has been shown to have advantages compared to others. It is on our future research agenda to build in various anonymization approaches. The Mondrian algorithm uses greedy recursive top-down partitioning of the (multidimensional) quasi-identifer 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.

3 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.

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. 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 then used the dataset for evaluating the accuracy of our attribute extraction component (discussed in Section 2.3).

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 2.4). We then evaluate the utility of the anonymized data through a set of queries.

Effectiveness of Attribute Extraction. To evaluate the effectiveness of our attribute extraction component, we conducted a set of experiments using 10-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).

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 are not reporting specificity because the non-identifying attributes are dominating compared to the identifying attributes so specificity will be always close to 100% which will not be very informative.

Table 2. Effectiveness of Attribute Extraction

.

Overall Accuracy: 0.982 (2675 / 2725)			
Label	Prec	Recall	F1
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

Table 2 presents the extraction results in precision, recall and $F1$ metric for each identifying attribute (class) as well as the overall accuracy. Most attributes achieve nearly perfect performance, but finding proper names (and how long those names extend) can be still further improved. 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 [19], part of speech tagging to further improve the performance for various datasets.

Effectiveness of 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 defined as % of correct reports being returned. Concretely, we randomly generated 10000 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.

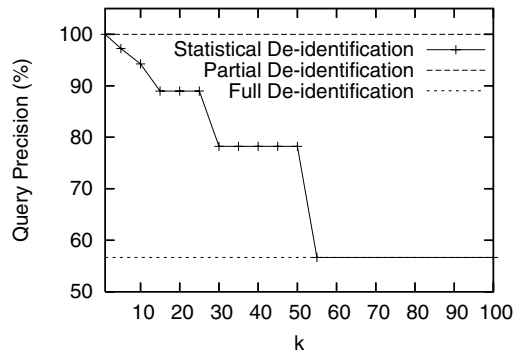

Figure 3. Effectiveness of De-Identification

Figure 3 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.

4 Conclusion and Future Works

We presented a conceptual framework as well as a prototype system for anonymizing 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 (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, in collaboration with the Ontology Informatics group at the Winship Cancer Institute at Emory, 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)[5] will also be carried out.

References

[1] C. C. Aggarwal. On k-anonymity and the curse of dimensionality. In *VLDB*, pages 901–909, 2005.

[2] G. Aggarwal, T. Feder, K. Kenthapadi, S. Khuller, R. Panigrahy, D. Thomas, and A. Zhu. Achieving anonymity via clustering. In *PODS*, pages 153–162, 2006.

[3] 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.

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

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

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

[7] 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.

[8] V. S. Iyengar. Transforming data to satisfy privacy constraints. In *KDD*, pages 279–288, 2002.

[9] D. Kifer and J. Gehrke. Injecting utility into anonymized datasets. In *SIGMOD Conference*, pages 217–228, 2006.

[10] 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.

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

[12] K. LeFevre, D. Dewitt, and R. Ramakrishnan. Incognito: Efficient full-domain k-anonymity. In *ACM SIGMOD International Conference on Management of Data*, 2005.

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

[14] K. LeFevre, D. DeWitt, and R. Ramakrishnan. Workload-aware anonymization. In *SIGKDD*, 2006.

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

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

[17] C. Manning and H. Schutze. *Foundations of Statistical Natural Language Processing*. MIT Press, 1999.

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

[19] A. McCallum. Efficiently inducing features of conditional random fields. In *19th Conference in Uncertainty in Articifical Intelligence (UAI)*, 2003.

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

[21] A. Meyerson and R. Williams. On the complexity of optimal k-anonymity. In *PODS*, pages 223–228, 2004.

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

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

[24] 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.

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

[26] L. Sweeney. Guaranteeing anonymity when sharing medical data, the datafly system. In *Proceedings of AMIA Annual Fall Symposium*, 1997.

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

[28] R. K. Taira, A. A. Bui, and H. Kangarloo. Identification of patient name references within medical documents using semantic selectional restrictions. In *AMIA*, 2002.

[29] S. M. Thomas, B. Mamlin, and G. S. adn C. McDonald. A successful technique for removing names in pathology reports. In *AMIA*, 2002.

[30] T. M. Truta and B. Vinay. Privacy protection: p-sensitive k-anonymity property. In *ICDE Workshops*, 2006.

[31] K. Wang and B. C. M. Fung. Anonymizing sequential releases. In *ACM SIGKDD*, 2006.

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

[33] X. Xiao and Y. Tao. Anatomy: Simple and effective privacy preservation. In *VLDB*, pages 139–150, 2006.

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

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

[36] S. Zhong, Z. Yang, and R. N. Wright. Privacy-enhancing k-anonymization of customer data. In *PODS*, 2005.

[5] Cancer Biomedical Informatics Grid. https://cabig.nci.nih.gov/