International Journal of Cyber-Security and Digital Forensics (IJCSDF) 1(1): 41-49
The Society of Digital Information and Wireless Communications, 2012 (ISSN: 2305-0012)

Privacy in Medical Data Publishing

Lila Ghemri and Raji Kannah

Department of Computer Science, Texas Southern University
3100 Cleburne Street, Houston, TX 77004
Ghemri_lx@tsu.edu
rajikannah@tsu.edu

ABSTRACT:

Privacy in data publishing concerns itself with the problem of releasing data to enable its study and analysis while protecting the privacy of the people or the subjects whose data is being released. The main motivation behind this work is the need to comply with HIPAA (Health Insurance Portability and Accountability Act) requirements on preserving patient's privacy before making their data public. In this work, we present a policy-aware system that detects HIPAA privacy rule violations in medical records in textual format and takes remedial steps to mask the attributes that cause the violation to make them HIPAA-compliant.

KEYWORDS

Data publishing, privacy, medical, HIPAA.

1. DATA PUBLISHING

Publishing public or semi public data and records offers tremendous benefits to many fields. It allows government agencies to predict and plans for future needs, it allows scientists to develop models of the data being observed, find patterns and connections between attributes and advance their respective fields to realize benefits for the whole humanity. Furthermore with the advent of the Web and the networked world, huge amounts of data are being stored in databases and becoming increasingly available to study, mine and analyze. These advances, however, do not come without a price. Indeed, making these data available, also uncovers vulnerabilities as people private information about themselves, their health, their shopping habits becomes public and the cozy anonymity of being a record amongst thousands suddenly singles out a specific person with their name, age, address all disclosed [1]. This paper is organized as follows: Section 2 presents an overview of the methods used for privacy protection of datasets. Section 3 talks about the laws and regulations in effect to protect medical data. In section 4, we present PACS, a Privacy Aware and System. Section 5 in presents the k-means algorithm and its use in testing the utility of our approach; we will also discuss the results and conclude in section 6.

2. PRIVACY IN DATA PUBLISHING

Privacy in data publishing concerns itself with the problem of releasing data to enable its study and analysis while protecting the privacy of the people or the subjects whose data is being released. Rastogi *et al* [2] give the following definition of privacy in data publishing: "given a database instance containing sensitive information, "anonymize" it to obtain a view such that on one hand attackers cannot learn any sensitive information from the view, and on the other hand legitimate users can use it to compute useful statistics". These two goals may seem to be opposite and contradictory, too little anonymization and the sensitive data can be reconstructed by the attacker,

too much anonymization and the data is no longer useful for researchers to analyze, this dilemma has been coined as Privacy Versus Utility. In determining what attributes and what values in the database need to be anonymized; designers first select the database attributes that can uniquely identify a record. These attributes such as names, addresses, SSN, are called the *identifiers* and are usually removed or blacked out from any view. An attacker will expect such values not to be available to him/her. The other attributes that are kept in the released view of the database are called *quasi-identifiers. (QI)*. The QI are usually anonymized, and an attacker will usually rely on their values together with other external sources to reconstruct the values of the identifiers. For example, assume that the table below is the Employee table of ATCO Company.

Table 1: Original Employee Data

Name	SSN	Income	Sex	Age
John Kean	123-45-6789	52,000	M	32
May Sim	234-56-7890	60,000	F	46
Paul Reve	345-67-8901	23,000	M	28
Corry Jay	456-78-9012	34,000	F	31

With Table 2, an adversary who knows that Corry, who is female, works for ATCO and that her age, is less than 30 can easily infer that Corry's salary is 34,000.

Table 2: View with Names and SSNs masked

Name	SSN	ID	Income	Sex	Age
(John Kean)	(123-45-6789)	1	52,000	M	38
(May Sim)	(234-56-7890)	2	60,000	F	44
(Paul Reve)	(345-67-8901)	3	23,000	M	31
(Corry Jay)	(456-78-9012)	4	34,000	F	28

Table 3: View with 4 attributes anonymized

Name	SSN	ID	Income	Sex	Age
(John Kean)	(123-45-6789)	1	52,000	Any	[35- 45]
(May Smith)	234-56-7890	2	60,000	Any	[35-45]
(Paul Reve)	(345-67-8901)	3	23,000	Any	[25-34]
(Corry Jay)	(456-78-9012)	4	34,000	Any	[25-34]

However, the same adversary can only guess Corry's salary with 50% probability using Table 3. So Corry's privacy is improved as more QI are anonymized and masked.

In order to insure that the data released preserves the privacy of the record and still contains some useful information, three components have to be taken into account, as defined by [3]:

Sanitization Mechanisms: Given an original dataset, a sanitization mechanism is a process or a program that makes the data less accurate or anonymizes it. This mechanism defines a space of possible views of the dataset. Each possible view is a release candidate. Masking is an example of such mechanisms.

The most common techniques for anonymization are:

- *Data Swapping:* Some selected attributes values are swapped between records, for example, the age of May and John may be swapped.
- *Randomization:* Random noise is added to the original data creating a sanitized view of the data. For example, adding 5 years to each age.
- *Generalization:* Consists of replacing the original value by a semantically consistent but less specific value, for example replacing Paul and Corry salaries with (< 35,000)
- *Suppression:* Replaces some attribute values by a special symbol such as "*" or "Any". Suppression can be considered as an extreme case of generalization.

Privacy Criterion: Given a release candidate, the privacy criterion defines whether the candidate is safe for release or

Utility Metric: Given a release candidate, the utility metric evaluates the utility of the release candidate, or the information loss due to the sanitization process.

3 PRIVACY IN MEDICAL DATA

Medical Data is the most widely sought after and used data for research purposes.

The goal of privacy preserving in publishing medical data is to protect the patient's confidential information from unwanted disclosure. Consequently, privacy protection policies of medical records have become law in the United States with HIPAA (American Health Insurance Portability and Accountability Act of 1996) [4]. HIPAA is a set of regulations with which doctors, hospitals and other health care providers have to comply. HIPAA seeks to ensure that all medical records, medical billing, and patient accounts meet certain consistent standards with regard to documentation, handling and privacy. In particular, HIPAA Privacy Rule provides federal protection for personal health information held by medical entities and gives patients an array of rights with respect to that information. At the same time, the Privacy Rule tries to keep balanced so that it permits the disclosure of personal health information needed for patient care and other important purposes.

HIPAA defines *Individually identifiable health information* as information that is (a) a subset of health information, including demographic information (b) collected from an individual by the healthcare organization and (c) relates to the past, present, or future physical or mental health or condition of an individual; the provision of health care to an individual; or the past, present, or future payment for the provision of health care to an individual; and (d) identifies the individual; or (e) with respect to which there is a reasonable basis to believe the information can be used to identify the individual. In particular, HIPAA specifies that all information related to an individual medical condition, treatment and procedures is individually identifiable health information and needs to be privacy-protected before its release [5]. Although many health organizations follow the privacy procedures outlined by HIPAA, the problem of privacy breaches and intentional or unintentional release of private information is a real one. According to the *Wall Street Journal*, between April of 2003 and November 2006, the Department of Human Health Services fielded 23,886 complaints related to medical-privacy rules, but it has not yet taken any enforcement actions against hospitals, doctors, insurers or anyone else for rule violations [6].

3.1 Medical Data De-identification

Clinical medical records contain patients' health information and are a prime source for researchers in the medical field to extract and mine. These records are in a free-form text format and they are usually processed using methods from natural language processing (NLP) such as Information Extraction, Parsing, etc. Most of the methods that apply in database mining can also be applied to text mining [7]. However, text presents its own challenges, such as the high dimensionality; in that every word becomes an attribute in the database, imprecise semantics: what does each attribute really mean, and uncertainty about which attributes are identifiers and/or quasi-identifiers [8], [9].

In our approach, we have focused on two quasi-identifiers that have been identified by HIPAA: drug names and drug dosages under the (reasonable) assumption that patients do not want their medical condition disclosed to people who are not directly providing health care for them. Knowing what type of medication a patient is taking, gives a very strong indication about their medical condition. For example knowing that a patient is taking Ativan, one can conclude that they may be suffering from depression.

Similarly, the dosage with which a drug is administrated, can uncover the patient's

condition, for example, the drug *Reclast* is usually used in lower doses to prevent osteoporosis but is used in higher doses to treat bone cancer.

3.2 Privacy Models for Medical Records

The K-anonymity model, introduced by Sweeney [10], [11], [12] protects against record linkage, in that it ensures that the database view that is released contains at least k records that hold similar values for the public data in the quasi-identifiers. In a linkage attack, it is assumed that the attacker has full knowledge of the public attribute values of individuals, but no knowledge of their private data. The attacker uses external public tables containing the identities of the individuals, such as voters' lists, and the public attributes contained in the QI. A linking attack is successful if the attacker is able to match the identity of an individual against the value of a private attribute.
In order to realize k anonymity, two conditions are required for a database T: (1) the set of all tuples in T containing identical values for the quasi-identifier set belongs to an equivalence class. (2) Every tuple is in an equivalence class of size at least k.
This model was further refined by [13] who noticed that even k records in a database do not necessarily have different values of their private data and suggested an approach called l-diversity, which ensures that every group of tuples that share the same QI values in the table, have at least *l* distinct sensitive values that are of roughly equal proportion. In an another work, Gil et al, presented a framework to protect patients' privacy in the medical data workflow [14]

4 PACS

The present work is an extension of our previous work in which lexical entailment was explored as a means for data de-identification [15]. The contributions of this paper are twofold. We first present a **P**rivacy **A**ware **C**orrecting **S**ystem (PACS) which encompasses detection of private information and remediation by suppression and generalization of the data. Then we focus on measuring the utility of our approach in the context of clustering and unsupervised learning.

4.1 Health Information Dataset:

The health information used in this work was contained in a database obtained from the Intensive Care Unit Safety Reporting System (ICUSRS). The ICUSRS was developed by a team of medical and public health researchers at the Johns Hopkins University. It collected data from 30 intensive care units (ICUs) across the United States about harmful or potentially harmful incidents in ICUs and to apply these data to improvement efforts in patient safety. The ICUSRS collects information about all types of incidents, i.e., events that could or did lead to patient harm, whether it involved a medication, device, fall, or other event [16]. Whereas most fields consisted of categorial values such as computer/software malfunction or computer/software error, etc, the database also included a field called Description in which the doctor or the nurse would explain the incident in their own words using plain language as in Figure 1.

"Patient on vasopressors noted to have dose of phenylephrine which was 10 times recommended maximum dose."

Figure 1. Example of Description Field

Figure1 is an example of the Description field in the database in which Quasi Identifiers (QI) such as drug names and dosages have been left intact and published.

4.2 Overview of PACS

Medical data is usually contained in databases with crisp attributes and semantics. However, most entries that are made by practitioners in the medical field are recorded as text, these are usually called doctor or nurses notes and are transcribed in the database as a long text fields and are released in public databases with very little modification. Our system focuses on the doctors' or nurses' notes and analyzes them. It is comprised of two main modules: The detection module and the de-identification module.

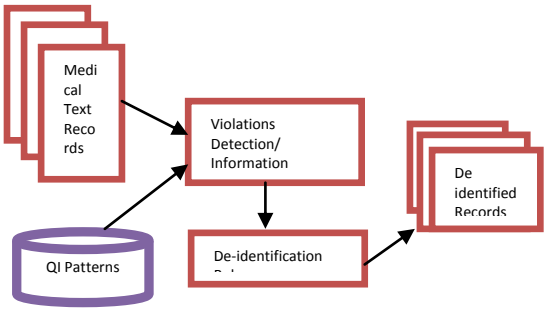

Figure 2. PACS organization

4.2.1 The Detection Module

This module scans the textual part of the medical record, in order to find patterns of drugs. If such a pattern is detected, then the record is flagged as needing sanitization and the record is passed to another module, called the de-identification module**.** The Detection Module task is one of Entity Recognition (ER) in which the entities looking for are drug names as well as drug dosages. In ER tasks, two main approaches are usually used: either a rule-based approach in which the program designer hand codes the patterns that denote the pattern they are searching for. This approach is very efficient and produces high yield, however it suffers from being labor intensive. The second approach is to use machine learning techniques to enable a program to learn the patterns. This technique requires a non-negligible overhead effort in annotating a corpus with the correct labels which will be used for training purposes. It also usually requires a sizable corpus to obtain valid learners. Since we were limited by the size of the corpus to only a couple of thousands records, we opted for the first method. We designed two recognizers: a drug name recognizer that used semantic patterns to detect the presence of a drug name, such as "Patient on <medication>" or "dose of <medication> ". Our recognizer has several such rules and has been implemented in Java. A grammar that recognizes drug dosages was also designed and implemented in Perl.

4.2.2 The De-identification Module

The purpose of this module is to apply generalization rules so as to mask the private information and make the record fit for release. This module receives a record which has been flagged with a violation and uses generalization rules to mask the offending data. We used the **Global Recoding principle** to generalize the attribute. **Global Recoding principle** is a generalization method for each QI attribute in which a hierarchy of concepts is used to generalize a QI attribute to its parent QI concept [17].

4.2.2.1 Global Recoding

A single dimensional global recoding for input data set D with QI attributes $Q1,...,Qn$ is defined by a family of n generalization functions φi: $dom(Qi) \rightarrow dom(Q\_i)$, such that the values in $dom(Q\_i)$ are semantically consistent generalizations of the values in $dom(Qi)$.

Levels of Generalization

- Let the j^{th} attribute have domain D^j and l^j levels of generalization. Let the partition corresponding to the h^{th} level of generalization be for $1 \leq h \leq l^j$.
- Let a value $y \in D^j$ when generalized to the h^{th} level be denoted by $g_h(y)$,
- A *generalization function* h is a function that maps a pair (i, j), $i \leq n$, $j \leq m$ to a level of generalization $h(i,j) \leq lj$.
- Semantically, h(i, j) denotes the level to which j^{th} component of the i^{th} vector (or the $(i,j)^{th}$ entry in the table) is generalized.
- Let h(xi) denote the *generalized* vector corresponding to x_i, i.e. $h(xi) = (g_h^{(i,1)}(x_i[1]), g_h^{(i,2)}(x_i[2]), g_h^{(i,m)}(x_i[m]))$.

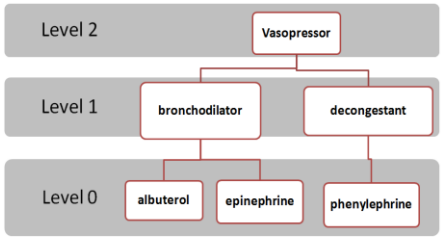

Figure 3. Hierarchy of QI concepts

- For each attribute QI, a generalization hierarchy exists.
- Each level of generalization corresponds to a partition of the attribute domain.
- A partition corresponding to a given level of the generalization hierarchy is a refinement of the partition corresponding to the next higher level.
- Singleton sets correspond to absence of generalization and correspond to the highest level of generalization.

For example, using the hierarchy described in Figure 3, applying the generalization rules at level 1, on the record shown in Figure 1, we obtain:

"Patient on vasopressors noted to have dose of decongestant which was 10 times recommended maximum dose."

Applying the generalization rules at level 2, will produce the following record:
"Patient on vasopressors noted to have dose of vasopressor which was 10 times recommended maximum dose."

However, PACS takes the generalization one step further and suppresses all drug names, in which case the record becomes:

"Patient on medication noted to have dose of medication which was 10 times recommended maximum dose."

We elected to generalize at this level because the names of the drugs involved are usually irrelevant in the description of the adverse event and this is more in tune with HIPAA.

5. UTILITY EVALUATION

The tradeoffs between privacy preservation and utility are always at the forefront of any system designed for that purpose. In order to measure the effect of the data modification rules that PACS applies to the data, we tested our approach on an unsupervised learning task and specifically a clustering task. After applying PACS, we obtained 3 "anonymized" datasets with varying degrees of data masking: One dataset with all dosages suppressed and all drug names left untouched. One dataset with all drug names (no D) replaced with the word "medication" and all dosages kept (no M). And a third dataset in which both the drug names and the dosages have been removed (no MD). We then applied a clustering algorithm to each dataset, including the original dataset and compared the clusters obtained.

We used RapidMiner™ as a tool to perform the text mining tasks [18]. We selected the K-means clustering algorithm and run the clustering program on the four datasets.

5.1 Record Representation

Classifiers and learning algorithms cannot directly process text documents in their original form. In order to reduce the complexity of documents and make them easier to handle, documents have to be transformed from the full textual version into a document vector. A document vector describes the contents of the document in terms of the words present in the document or features and a weight associated with each feature. Consequently, each word present in the document represents a feature, and its weight is usually a function of frequency of occurrence of that given word in the document. During the preprocessing phase, each document is represented as a feature vector in this space (sequence of features and their weights), the whole data set is represented as a matrix of n rows with m attributes, consisting of n, m-dimensional vectors: $x_1 \ldots x_n$.
Each document $x_i=(v_0, v_1, v_2, \ldots\ldots, v_m)$ such that:

$$v_i \begin{cases} =1 & \text{if word } w_i \text{ is present in document } x_i \\ =0 & \text{otherwise} \end{cases}$$

Furthermore, in order to focus on the most salient features, several pre-processing steps have to be performed in sequence on the text before mining is possible. These are (1) Tokenize the string of words which breaks the input into distinct words. (2) Filter the most common English words, using a predefined list. (3) Stem words into their root form using the Porter stemmer, (4) Remove words of length less than 3 letters.

These steps are depicted in Figure 4.

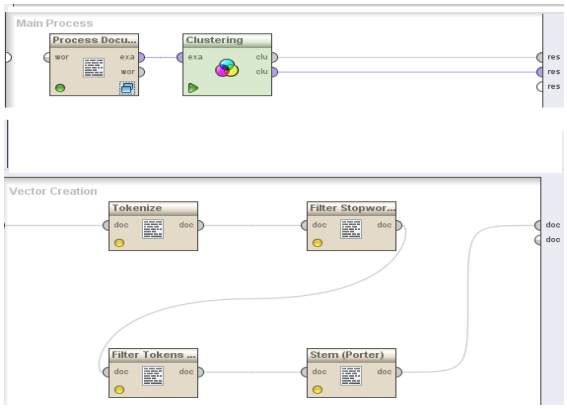

Figure 4. Preprocessing steps

After the preprocessing phase are performed, the K-means algorithm is then applied as a clustering method with k=5. This effectively clusters our dataset into 5 clusters.

5.2 The K-means Clustering Method:

K-means is a clustering algorithm that is widely used [19]. It partitions a collection of n vectors into a set of k ($k \leq n$) clusters $\{c1, c_{2\ldots} c_k\}$ in which each vector belongs to the cluster with the nearest mean. The flowchart of the algorithm is represented below:

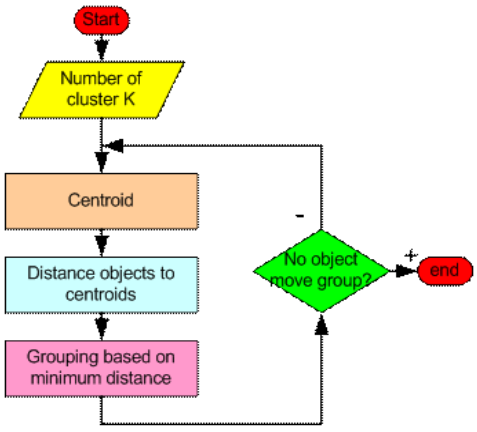

Figure 5. Flow chart of K-means algorithm ([19])

5.3 Utility of Disclosure

In order to quantify the data and clustering precision loss, we compared the K-means clustering results of the modified datasets to the clusters obtained from the original dataset and computed the number of records that stayed in their original cluster, we call this measure "*persistence"*. If the *persistence* is high, then the de-identification rules are not drastically compromising the mining process. If the persistence is low, then the anonymization rules have caused records to cluster in different groups and changed the outcome of the original clustering process, thus lowering the utility. We calculated the distance between two clusters C_i and C_j as the difference between the union set of the two clusters and the intersection set of the two clusters. If C_i and C_j were identical, then the distance is equal to 0:

We normalized the value obtained above and expressed as a percentage: *persistence*. The results are presented in the charts below:

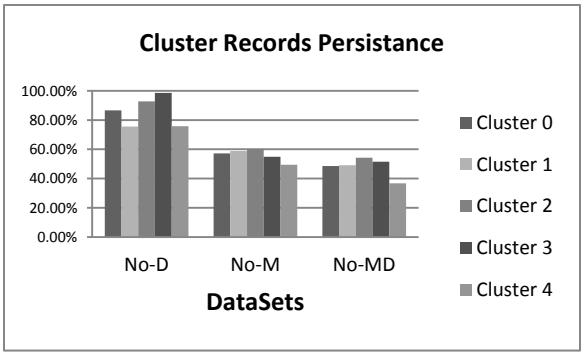

Figure 6. Record/Cluster Persistence

Charts in Figure 6 and 7 show that the clustering results degrade slowly as more data gets generalized by PACS. It also shows that there is good persistence with the no-dosage dataset (98.6% to 74.5%), average 86%, then the persistence somewhat decreases for the second dataset (with drug names masked) (60%-50%), average: 55%, whereas it strongly degrades for the dataset with both dosages and drug name removed (51%-39%), average 45%

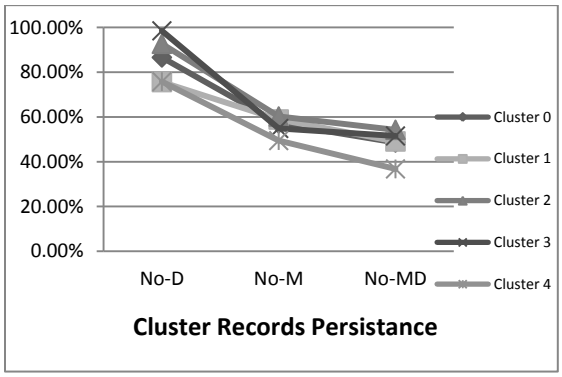

Figure 7. Persistence Graph

5.4 Disclosure Risk versus Utility of Disclosure:

In the paragraph above, we discussed the utility of disclosure of the sanitized versions of the dataset. In this part, we will present the disclosure risk of the release. The disclosure risk is defined as the product of the fraction of the original record release with the percentage of sanitized records which still cluster together or persistence [3]. To compute these numbers, we randomly sampled 1% of our dataset, then 5% of the dataset, then 10% of dataset and used the average persistence numbers computed above; the results are in Table 5.

Table 4. Disclosure Risk

Dataset	1%	5%	10%
No D	.86%	4.5%	8.6%
No M	.55%	2.75%	5.5%
No MD	.45%	2.25%	4.6%

Table 4 shows that the disclosure risk is greatest if only the dosage attribute is masked and as larger portions of the data is released. However the risk decreases as more data is masked and less of the dataset is published.

6. CONCLUSION:

In this work, we presented a privacy-aware system that detects HIPAA privacy violations in doctors' or nurses' notes using semantic patterns and grammars. Our system masks the patients' private information by applying concept generalization. We tested our approach by anonymizing a dataset of medical records and obtained three datasets with varying degrees of anonymization. We then run a clustering algorithm on the sanitized datasets and measured record to cluster persistence. We also mitigated the results by computing the disclosure risk associated with publishing portions of each dataset. We plan on expanding this work to use machine learning techniques, so that larger datasets can be processed.

7. REFERENCES

1. Narayanan, A., Shmatikov, V.: De-anonymizing Social Networks: SP'09 Proceedings of the 2009 30th IEEE Symposium on Security and Privacy, 173--187 (2009)
2. Rastogi, V., Suciu, D., Hong, S.: The Boundary Between Privacy and Utility in Data Publishing : Proceedings of the 33rd International conference on Very large data bases, 532--542 (2007)
3. Chen, B –C., Kifer, D., LeFevre, K., and Machanavajjhala, A.: Privacy-Preserving Data Publishing, Foundations and Trends in Databases Vol. 2, 1--167(2009)
4. Summary of the HIPAA Privacy Rule http://www.hhs.gov/ocr/privacy/hipaa/understanding/summary/index.html
5. HIPAA 101**:** HIPAA Privacy Rule and Compliance, http://www.hipaa-101.com/hipaa-privacy.htm
6. Health Insurance Portability and Accountability Act, http:// wikipedia.org/wiki/Health_Insurance_Potability_and_Accountability_Act
7. Zhou, X., Han, H., Chankai, I., Prestrud, A., and Brooks, A.: Approaches to Text Mining for Clinical Medical Records. In: Haddad, H. (ed.) In: Proc 2006 ACM Symposium on Applied Computing (SAC) pp 235--239. Dijon, France, (2006).
8. Aggarwal, C., and Yu, P.S. (eds.): Privacy Preserving Data Mining: Models and Algorithms. Springer, (2008)
9. Aggarwal, C.: A General Survey of Privacy-Preserving Data Mining Models and Algorithms: Privacy Preserving Data Mining: Models and Algorithms. In Aggarwal, C., and Yu, P.S. (eds.) pp. 11--52 Springer, (2008)
10. Sweeney, L.: Achieving K-anonymity privacy protection using generalization and suppression. In: International Journal on Uncertainty, Fuzziness and Knowledge-based Systems, 10 (5) pp 571--588 (2002)
11. Sweeney, L.: Guaranteeing anonymity when sharing medical data, the datafly system. In: Proc 1997 AMIA Annual Fall Symposium, pp. 51--55 Washington DC (1997).
12. Sweeney, L.: Replacing Personally-Identifying Information in Medical Records, the Scrub System. In: Proc 1996 AMIA Annual Fall Symposium, pp. 333--337 Washington (1996).
13. Li, N., Li, T., Venkatasubramanian, S.: *t*-Closeness: Privacy Beyond *k*-Anonymity and *l*-Diversity. In: Proc IEEE International Conference on Data Engineering (ICDE), pp. 106-115 (2007).
14. Gil, Y., Cheung, W.K., Ratnakar, V., , Chan, K-K: Privacy Enforcement in Data Analysis Workflows In: Proc 2007 Workshop on Privacy Enforcement and Accountability with Semantics, South Korea , pp. (2007) .
15. Ghemri, L., Kannah, R.: Using Lexical Entailment for Privacy Protection in Medical Records. In: Proc. 2012 International Conference on Informatics and its Applications, pp.228--232 Malaysia (2012)
16. Holzmueller, C. *et al* .: Creating the Web-based Intensive Care Unit Safety Reporting System: Journal Am Med Inform Assoc.12 pp.130-139(2005)
17. Gardner, J.: Privacy Preserving Medical Data Publishing. PhD dissertation, Emory University (2012)
18. RapidMiner™ , www.rapidminer.com/
19. *k*-means clustering, http://en.wikipedia.org/wiki/K-means_clustering.