ICEHTM 2014

The 2nd International Conference on e-Health and Telemedicine

www.icehtm.net

Mary Ann Liebert, Inc. publishers

DIAGNOSTIC PATHOLOGY

Role-based privacy-preserving health records distribution

Pelin Aktaş, Hayri Sever, Murat Aydos
Department of Computer Engineering, Hacettepe University
{pelinaktas, sever, maydos}@hacettepe.edu.tr
corresponding author: pelinaktas@hacettepe.edu.tr

ABSTRACT

The data obtained by health organizations give numerous opportunities for generating solutions ahead. It is essential that the accurate data are shared in order to get useful results within healthcare systems. Accurate records of personal health data include sensitive information about individuals. Hence sharing the subject records bearing on original structure paves the way for disclosure of personal privacy.

In recent years, Privacy-Preserving Data Mining (PPDM) and Privacy-Preserving Data Publishing (PPDP) approaches have been extensively studied in order to protect personal privacy and security. In this study, different approaches to PPDM and PPDP are summarized and evaluated within the framework of health records. In line with evaluation phase conducted, we propose three general role-based system architectures, collecting health data from health organizations, satisfying anonymization on the collected data and publishing health information securely among several parties. In the anonymization phase, common methods (k-anonymity and ℓ-diversity) are applied on the collected data in order to compare the architectures. In the publication phase, a role-based publishing approach is implemented, in which only the required partition of personal health records are released for a specific organization by using database partitioning methods. The purpose of the proposed architectures is to achieve the desired balance on the trade-off between the levels of privacy protection and data utility.

Keywords: privacy-preserving, anonymization, database partitioning, distribution, personal health records, generalization

1. INTRODUCTON

The data obtained by many organizations and corporations pave the way for generating great solutions on any field. Taken into consideration using such data on the process of information extraction, many new opportunities will arise. Specially, in healthcare systems, old data about patients will give many opportunities for taking measures to prevent diseases in advance. However, such data about individuals contain sensitive information, and sharing or publishing this data records bearing on original structure violates personal privacy. Most privacy concerns are enhanced in the context of legal perspective rather than technical conception especially in the field of health-care information like U. S. Department of Health And Human Services, Office for Civil Rights [HIPAA]. This study and a lot of work, on the concern of privacy-preserving data, are interested in measuring privacy.

In recent years, Privacy-Preserving Data Mining (PPDM) and Privacy-Preserving Data Publishing (PPDP) approaches have been extensively studied in order to protect personal privacy and security. Additionally the problem has been endeavored to solve in many research communities as well, such as the statistical disclosure community, cryptography community and database community. PPDM approaches have many models and algorithms for protecting personal or corporative privacy. The PPDM has aim of extracting relevant knowledge from large amounts of data with traditional data mining techniques while protecting sensitive information at the same time. PPDP focuses publishing data record about individuals (i.e., micro data) not data mining result. PPDP anonymizes the data by hiding the identities of record owners. While statistical community works on privacy-preserving publishing methods for statistical tables, most recent works in this field focus on Statistical Disclosure Control (SDC) methods. Database community generally works with other communities for protecting sensitive data using different methods such as database partitioning and query auditing. While cryptography community pays attention to common function for different companies to share their data securely without disclosure of sensitive data, most recent works on this field focus on Secure Multipart Computation (SMC). In some cases, the parallel lines of the works are quite similar, but the communities are not sufficiently integrated for the provision of broader perspective. Detailed information was given Section 1.

1.1. Contributions

In this study, several privacy-preserving approaches are summarized and evaluated within the framework of health records. Same techniques in these approaches do not preserve the truthfulness of values at the record level such as randomization in order to indicate mining results or securely distribution not re-identification. In that case for health records, the released data become nonessential and useless. Therefore in this study, we focus on selected approaches for anonymization and publishing techniques according to truthfulness of values. In line with evaluation phase conducted, we propose three role-based system architectures, collecting health data from health organizations, satisfying anonymization on the collected data and then publishing it among organizations interested in personal data. We mainly focus on architectural issues rather than mining the data or providing security on publication phase. Our research consists of three main contributions in three different scenarios. In a role-based health data distribution approach, only the required

partition of personal health data will be released for a specific organization by using database partitioning methods. We illustrate three scenarios for distributing personal health records among several parties. The first scenario in which parties anonymize their dataset according to given criteria and send it to the central database, has architecture of untrusted third party approach. The second scenario, in which parties send their data to the central database, has architecture of trusted third party; it is out of concern which protocol they use. Anonymization is applied to the collected data in central database with determined criteria and then this anonymized data are partially published among parties. The last scenario has architecture of trusted third party, but implements different methods in publication. In this scenario, the collected data in the central database are firstly partitioned according to the requirement of parties and each partition is anonymized before distribution. Each scenario has its pros and cons detailed in Section 5.

The purpose of the proposed architectures is to achieve the desired balance on the trade-off between the levels of privacy protection and data utility.

In section 2, proposed architectures and background information are presented. In section 3, common threats in anonymization and privacy techniques across these threats are explained. Section 4 deals with the data partitioning and data distribution approaches. Section 5 covers the proposed role-based architectures. Section 6 includes experimental evaluations and finally section 7 concludes from the results of experiments and makes suggestion to the future works for further progress.

2. RELATED WORKS AND BACKGROUND

Our research is motivated by inspiration and information from a number of relevant areas. We discuss them briefly below.

2.1. Data Collection and Data Distribution

In Figure 1 [Fung B. et al. 2010], a basic scenario of data collection and publishing was described. In the data collection phase, the data holder collects data from record owners or organizations. In the data publishing phase, the data holder releases the collected data to the recipients in order to conduct an analyzing or mining process on the published data, but the real scenario has to be more complicated through the privacy. There must be privacy-preserving layer to protect record owners' sensitivity.

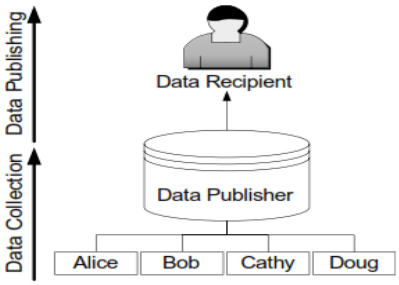


Figure 1: Data collection and data publishing [Fung B. et al. 2010]

In publication phase, it was assumed that two frameworks have trusted third party, the data publisher is trustworthy and record owners or organizations having records are willing to provide their information to the data publisher by ensuring that privacy will be protected. How the data collected from the data owners or organizations is out of this study. Any collected scheme can be used in the frameworks proposed in our study.

2.2. Privacy-Preserving Data Mining

A number of algorithmic techniques have been evaluated for privacy-preserving data mining. In [Aggrawal C.C. et al. 2008], recent models and algorithms are evaluated and parted according to their interested field and compared with different approaches. Data mining techniques in privacy-protection generally have been proposed for modifying or transforming the data. In privacy preserving data mining approaches, privacy of record owners is generally preserved by distorting the data values [Agrawal R. and Srikant R. 2000]. Since information on the distribution of the random data is used in order to distort the data, distorted data does not disclose identification of records and is useable for data mining safely. This study used random perturbation by adding noise to generate an approximation to the original data distribution. However, an attacker could filter the random noise, and thus violate privacy shown in [Huang Z. et al. 2005]. Because of that it contains data which do not exist in original database; randomly perturbed data is not truthful. In healthcare systems, the truthfulness of data has critical importance, so getting exact data rather than the approximate data is focused on in this study.

2.3. Privacy-Preserving Data Publishing

Privacy-Preserving Data Publishing (PPDP) approaches are focused on publishing the data, not mining or analyzing results. One of the aims of PPDP is to anonymize the data by hiding the record owner's identity. The work is generally interested in inference of sensitive attributes, background attacks, and measure of privacy with various information metrics. PPDP uses different data transformation techniques in order to perform the privacy preservation such as randomization [Agrawal R. and Srikant R. 2000], k-anonymity [Samarati P. and Sweeney L. 1998] and ℓ-diversity [Machanavajjhala A. et al. 2006]. In recent years, most PPDP techniques are evaluated to use published anonymous data for data mining

techniques or for making it usable for statistical community. In this study, privacy-preserving data publishing plays an important role in sharing personal healthcare information with different parties.

2.4. Anonymization Approaches

Most of the studies use anonymization to provide Privacy-Preserving systems in order to utilize the personal records. Anonymity means that no one, except from authorized user, cannot identify the certain records belongs to any specific individual. In [Samarati P. and Sweeney L. 1998], it was addressed the problem of releasing person-specific data while protecting the anonymity of the individuals to whom the data refer. In the work, it was illustrated that data holders remove all explicit identifiers such as name, address and phone number, from data so that other data, known as quasi-identifier such as birth date, gender and ZIP code in combination, can re-identify individuals. The work supports k-anonymization defined as each record is indistinguishable from at least k-1 other records with respect to certain explicit identifiers and quasi-identifiers attributes. The approach indicates how k-anonymity can be provided by using generalization and suppression techniques. However, k-anonymity cannot prevent sensitive attribute disclosure. An alternative, ℓ-diversity [Machanavajjhala A. et al. 2006], has been proposed as a solution to the problem, known as each set of rows corresponding to the same value for identifiers, contains at least ℓ well-represent values for each sensitive attribute. In Section 3, k-anonymity and ℓ-diversity approaches were detailed.

2.5. High Dimensionality and Data Partitioning

In anonymization approach, usually data sets have multi-dimension structure and while the dimension of data increases, the cost of computation also increases. Anonymization methods remain weak across high-dimensionality because of great computational cost. In [Meyerson A. et al. 2004], it is shown that optimal k-anonymization is NP-Hard. The study [Aggrawal C.C. 2005] discussed the difficulty of anonymization of the high-dimensional datasets and what kind of attacks can be enforceable on the datasets.

Data partitioning approaches are usually applied in the techniques of information sharing, privacy and security by using cryptographic methods [Pinkas B. 2002]. Sharing information among different parties needs security by the use of cryptographic protocols [Clifton C. et al. 2002]. This requirement is out of the present study. In the study, it was focused on how data come from different data collectors to different data recipients. In the literature, three partitioning ways used for sharing data includes horizontally partitioning [Lindell Y. and Pinkas B. 2000] which has the same attribute from different record owners, vertically partitioning [Dwork C. and Nissim K. 2004] which has the different attributes from the same record owners and hybrid partitioning which includes horizontally and vertically partitioning methods. In this study, vertically partitioning is used according to different party's requirements and horizontally partitioning is used to decrease high-dimensionality while hybrid partitioning is used to increase data utility.

3. THREATS IN ANONYMIZATION AND PRIVACY

Privacy-preserving approaches aim to build a system publishing data even in most adversarial environment, to protect sensitive information in published data and to keep data utility in maximal level. The fundamental method of these approaches is anonymization. In anonymization methods, it is assumed that data publisher has a table T which includes several subsets of attributes consisting of; *explicit identifiers (I)* containing information that explicitly identifies record owner and is typically removed from the released data such as name, social security number and cell-phone number, *quasi identifier (QID)* containing information that could potentially identifies record owner and is typically suppressed in the released data such as date-of-birth, gender and ZIP code, *sensitive attribute (S)* containing sensitive information about data owner and should be protected such as salary or disease, and *non-sensitive attribute* which does not fall into the previous three categories and can be published as it is when needed.

In [Sweeney L. 2002], it is shown that a person can be identified by using quasi identifiers even all explicit identifiers were extracted. While just one of the quasi identifier does not identify a record, combining with other attributes can disclose a person or a group. This case is known as Linking Attack in the literature. To prevent from this kind of attacks, the data publisher must provide an anonymous table T'.

Privacy models are divided into two categories due to the attack principles [Fung B. et al. 2010]. The first category shows an alteration according to attacker's background knowledge to link up the records. There are three types of linkage methods in this perspective. In a published data table, attacker is able to link a record owner to a record known as *record linkage*, to a sensitive attribute known as *attribute linkage*, to the published data table by itself known as *table linkage*. In all three types, it is assumed that the attacker knows record owner's quasi identifiers. The second category is based on uninformative principle [Machanavajjhala A. et al. 2006]. The published table should provide the attacker additional information about individual beyond the background knowledge without necessary linking it to a specific item in a dataset. If the attacker has a large variation between the prior and posterior beliefs, it is a *probabilistic attack*. Variety of attacks and their privacy models are described in [Fung B. et al. 2010]. Multiple privacy criteria have been proposed to prevent these types of attacks. Common criteria of anonymization, k-anonymity and ℓ-diversity are determined below and used in the present frameworks.

3.1. K-Anonymity

Anonymity is the most wide-spread approach in privacy protection systems. It aims at protecting datasets from identity disclosure which means that an adversary can learn sensitive information about individual by linking to a specific form of data item. In order to reduce the risk of identification, the k-anonymity technique was commonly proposed. In k-anonymity techniques [Sweeney L. 2002], granularity of the representation of quasi-identifiers was reduced by using generalization and suppression techniques. In generalization technique, the attribute values were generalized to a range in order to reduce specification. In suppression technique, the value of attribute was completely removed. Example of original data and anonymize data is shown in Figure 2. The k-anonymity approach required that each record, used in

literature as *tuple*, is indistinguishable at least k-1 other records with respect to quasi-identifier. Each group, which has at least k tuples indistinguishable from each other, forms an *equivalence class*. An anonymous table T' must consist of equivalence classes.

Age	Gender	Disease
(25-30]	Male	HIV
(25-30]	*	Fever
(30-40]	Male	Cancer
(30-40]	Female	Cancer
(30-40]	*	HIV

Age	Gender	Disease
29	Male	HIV
26	Female	Fever
30	Male	Cancer
32	Female	Cancer
30	Female	HIV

Figure 2: Original Data and Anonymized Data

3.2. ℓ-Diversity

The ℓ-diversity technique [Machanavajjhala A. et al. 2006] was proposed to prevent attribute disclosure which means that an adversary is able to infer additional information about an individual without necessary linking it to a specific item in a dataset. ℓ-diversity technique was generated after an observation that k-anonymity can create groups which leak information due to the lack of diversity in the sensitive attribute. The ℓ-diversity technique requires that every equivalence class has to contain at least ℓ "well- represented" sensitive values. ℓ "well-represented" means to ensure that there are at least ℓ distinct values for the sensitive attribute in each QID group. Different variants of ℓ-diversity have been proposed to implement different measures of privacy such as entropy-ℓ-diversity or recursive-(c, ℓ)-diversity.

Beside k-anonymity and ℓ-diversity, anonymization base methods, many alternative and stronger methods have been proposed in time such as t-closeness [Li N. and et al. 2007], δ-presence [Nergiz M. et al. 2007], ε-differential privacy [Dwork C. 2006]. Because of that this study focuses on frameworks which aim at sharing personal health information, achieves of anonymization methods were not detailed and used more.

4. DATA PARTITIONING AND DISTRIBUTION

Many cases on distributed privacy-preserving data mining wanted to obtain aggregate results from data sets which are partitioned. Partitioning could be horizontal, which has records distributed across multiple entities, or vertical, which has the attributes distributed across multiple entities. The main aim in most distributed methods was to allow computation of useful aggregate statistics over the entire data sets within the different participants. The problem of distributed data mining generally works with cryptographic field for determining Secure Multi-party Computation (SMC) [Pinkas B. 2002]. In our study, data partitioning methods were used for both data senders and recipients. For nation-wide benefit, especially in healthcare system, it has critical importance to collect data from organizations and to evaluate them in order to see potential threats. Therefore, the proposed architectures do not include only senders but also some statistical organizations. The collected data must be partitioned in order to send the collected data to different parties with maximum data utility. In this study, common computation methods were not detailed because this situation may change according to recipient party's requirements.

4.1. Distributed Algorithms for k-Anonymity

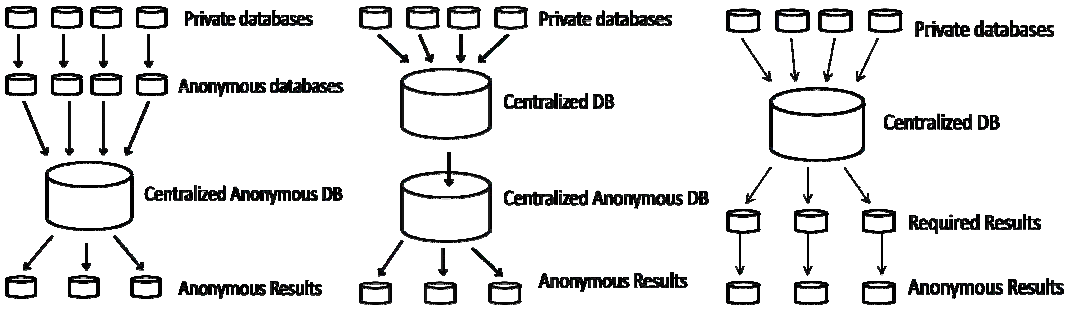
It is important to apply anonymity across different parties. In [Zhong S. et al. 2005], horizontally partitioned data have been discussed to maintain k-anonymity. The work tackled different situations in which each site is a costumer who has one record in the data. It was assumed that the data record has both sensitive attribute and QID attributes, and the sensitive attribute is encrypted until at least k records had the same values on the QID attributes. K-anonymous protocol was described in [Jiang W. and Clifton C. 2005] with vertically partitioned data across two parties. Common works include two parties which agree on the same QID to generalize the same values before publishing. In [Wang K. et al. 2005], it was discussed how the generalization is to perform accordingly two parties agreement before releasing, so k-anonymity was provided.

5. ROLE-BASED ARCHITECTURES

There are a number of potential approaches for privacy preserving data publishing and each of them may be applied to distributed databases. In this section, three scenarios were determined by using privacy-preserving publication or distribution frameworks.

5.1. Scenario-1

A simple approach is for each data collector to perform anonymization independently on their database for untrusted third party, shown in Figure 3(a). In this approach, each data sender anonymizes their data according to given anonymization criteria such as k-anonymity and ℓ-diversity. Then, the data sender sends this anonymized data to the central database. The central database integrates all anonymized data to obtain general situation and then generates different tables according to the receiver parties' requirements by database partitioning both vertical and horizontal. In the central database, each table is generated from integrated anonymized data. In this scenario, for data sender, a secure structure is provided, but for data receiver, minimal information utility is provided.



(a) Architecture of Scenario-1 (b) Architecture of Scenario-2 (c) Architecture of Scenario-3

Figure 3: Architectures of Scenarios

5.2. Scenario-2

In this scenario; data collectors send their data to the central database using any secure transmission protocols, shown in Figure 3(b). In this schema, it is assumed that central database is trusted third party and the data senders are willing to share their data for public utility. The original data which are collected in the central database are integrated then anonymized according to the given anonymization criteria. On the anonymized data, different tables are generated according to the requirement of receiver parties by using database partitioning methods. This scenario has provided more information utility compared to first scenario but it is considered insufficient. However, in this scenario, data tables are protected from some attacks such as table linkage. Since central database that generates receivable tables is anonymous, the published tables are anonymous as is in the central anonymous database. Therefore, integration of published tables does not disclose any privacy. However, this scenario is not appropriate for multi-dimensional databases because of its excessive computational cost.

5.3. Scenario-3

The last scenario proposed in this study has maximum data utility compared to the other two scenarios, shown in Figure 3(c). In this scenario, data collectors send their data to central database using any secure transmission protocols. In this schema, it is assumed that central database is trusted third party and the data senders are willing to share their data for public utility. The original data which are collected in the central database are integrated. After integration of all original data that come from data senders, different tables are generated according to requirements of receiver parties by using database partitioning methods. The tables which have original data are then anonymized according to given criteria and anonymous tables will become distributable to receiver parties. Using this scenario, data utility is increased but some attacks like table linkage is enforceable to distributed tables. Consequently, to protect privacy and conserve information utility, anonymization criteria of protecting table linkage such as δ-presence must be considered to adhere to the anonymization phase.

In all scenarios, it was assumed that the databases sent from data senders are homogeneous which means that all sites have the same schema, but each site has information on different entities. It was assumed that data senders are healthcare systems which have personal heath records and data receivers might not be in the health systems or semi-interested in healthcare information about individual, in all scenarios.

6. EXPERIMENTAL EVALUATIONS

6.1. Anonymization Metric

Privacy preserving methods have to take into consideration of the information retaining after processes. Anonymization problem is to produce an anonymous table T' that satisfies a given privacy requirement determined by the chosen privacy model and to retain as much data utility as possible. Two types of metric used for anonymization methods; *data metric* measures the data quality in the anonymous table T' with respect to the data quality in the original table T and *search metric* checks the parts of anonymization algorithms to provide an anonymous table T' with maximum information or minimum distortion. The two types of metrics usually share the same principle of measuring data quality.

In this study, *Discernibility Metric* (DM), proposed by [Bayardo R. et al. 2005], was used to measure utility in anonymization. DM endeavors to minimize the average equivalence class (E) size due to the fact that the more records are in the same equivalence class, the less specific information is preserved for these records. This data metric tries to make records indistinguishable with respect to QID in the entire database (D). The metric could be mathematically stated as follows:

6.2. Generalization Criteria

In this study, we performed a full-domain generalization [LeFevre K. et al. 2005] in anonymization which means that the same generalization techniques are applied to all values of QID. The technique used for generalization is Lattice and

Predictive Tagging which was illustrated into *Flash* algorithm [Kohlmayer F. et al. 2012]. The technique uses graduated generalization hierarchies known as *Generalization Lattice*. The example of generalization lattice of age and gender is shown in Figure 6. An arrow denotes direct generalization from specialized state. The state in the Figure 6, minimum generalization is "0,0" and maximum suppression is "5, 1" for these two attributes.

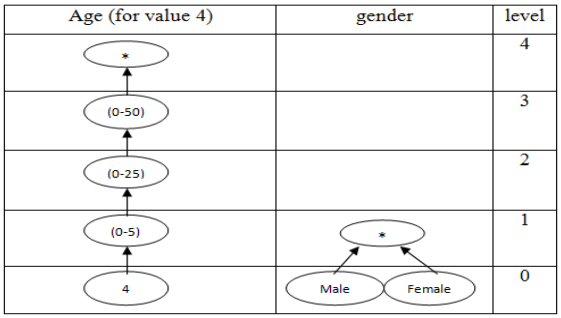


Figure 4: Generalization lattice of age and gender attributes

6.3. Dataset and Setup

The 2004 US census database (ADULT) [Newman D. et al. 1998] from UC Irvine Machine Learning Repository was used as main database. The dataset was modified to get rid of undefined attributes by removing record which had undefined attribute value. After modification, 29967 records and 7 attributes were remained to use. Three scenarios explained before have been implemented in Java within the open source data anonymization framework, ARX Data Anonymization Tool [Kohlmayer F. et al. 2012]. ADULT dataset was used to make different sender parties by horizontally random partition. Generalization lattice for full domain generalization schema was implemented by user-defined. For experiments, in each scenario, k-anonymization techniques were used with different k values which are 3, 4, 5 and 6. Additionally, ℓ-diversity technique was used in each scenario, but just 2-diversity was implemented because of used database's sensitive attribute which has only two different values. Each scenario was evaluated with monotonic discernibility metric.

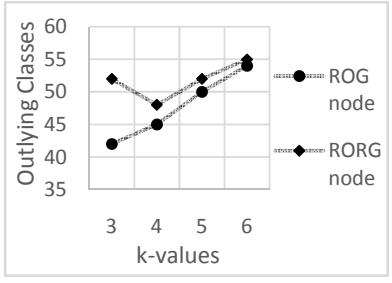
For evaluation of the scenarios, some common modifications were determined. Maximum suppression was determined as "4, 3, 3, 3, 1, 3" for 6 quasi-identifier attributes (age, work-class, education-num, occupation, gender, hours-per-week), and "1" for sensitive attribute (census-income). According to determined generalization hierarchy, the optimum generalization node is selected for anonymization process. In experiments, two different generalization nodes were used. The first node was optimum generalization node selected by metrics to obtain minimum information loss after anonymization. The second node was determined by user to obtain required generalization result. In experiments, the Result of Optimum Generalization node was shown as ROG node, and the Result of Optimum Required Generalization node was shown as RORG node.

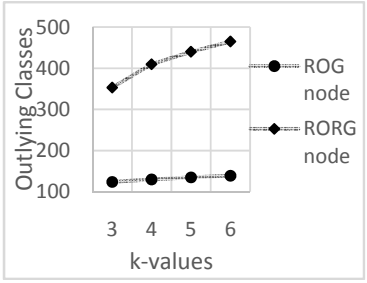
6.4. Results

Due to the monotonic discernibility metric, changing the k-values did not affect information loss or equivalence class number, it affected the number of outlying classes considered as exceptions and suppressed in anonymization.

In scenario 1, the used database was divided into three equal parts and it was assumed that each part belongs to different data senders. One of the divided parts was anonymized with different k-values and 2-diversity. In this scenario, required minimum generalization node was "3,2,2,2,0,2" for less suppression representation in anonymized table. The ROG node in this scenario was "2,3,3,1,0,2" for all k-values, and the RORG node was selected as "3,0,2,2,0,2" because it was optimum node in order to provide required generalization. The difference between used nodes and outlying classes for one sender party is shown in Figure 5.

In scenario 2, the all of the database was anonymized with different k-values and 2-diversity. In this scenario, required minimum generalization node was "3,2,2,2,0,2" for less suppression representation in anonymized table. The ROG node in this scenario was "1,3,3,0,1,3" for all k-values, and the RORG node was selected as "2,2,2,0,0,2" because it was optimum node in order to provide required generalization. The difference between used nodes and outlying classes this scenario is shown in Figure 6.





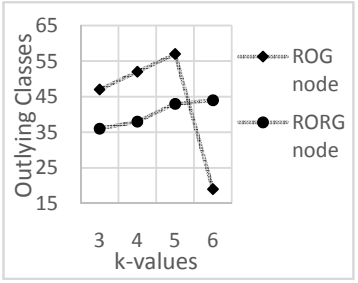


Figure 5: Outlying classes of Scenario-1

Figure 6: Outlying classes of Scenario-2

Figure 7: Outlying classes of Scenario-3

In scenario 3, the used database was partitioned vertically in order to select required attributes for anonymization. 4 quasi-identifiers (work-class, education-num, occupation and hours-per-week) were used for evaluation of this scenario. The partitioned part was anonymized with different k-values and 2-diversity. In this scenario, required minimum generalization node was "2,2,2,2" for less suppression representation in anonymized table. The ROG node in this scenario was "2,0,1,3" for the first three k-values, and the RORG node was selected as "1,2,0,2" because it was optimum node in order to provide required generalization. For the last k-value, because of that there were not any group providing ROG node, the ROG node turned into "3,0,1,3". The difference between used nodes and outlying classes is shown in Figure 7.

Considering the nodes, each scenario was divided into 2 parts for easy use. The first part of scenarios was evaluated with ROG node, shown as "1-a, 2-a, 3-a". The second part of scenarios was evaluated with RORG node, shown as "1-b, 2-b, 3-b".

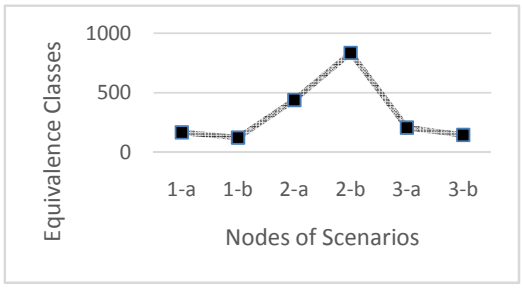


Figure 8: Equivalence classes distribution of Scenarios

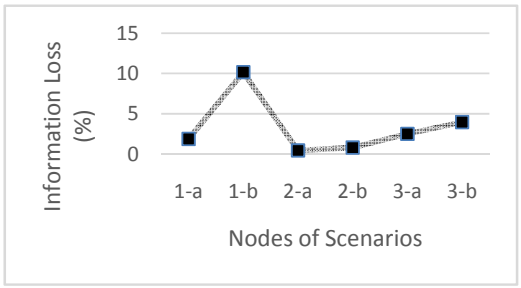


Figure 9: Information loss distribution of Scenarios

The ideal consequent results were obtained from 4-anonymity and 5-anonymity process according to the results of outlying classes in all scenarios due to the database. Figure 8 shows that the results of the distribution of equivalence class numbers obtained in 5-anonymization for all scenarios and nodes. The results of "1-a" and "1-b" show just one sender party's anonymization results but it would be around three times higher in all evaluations for entire database. In figure 9, the information loss of all scenarios and nodes in 5-anonymity is showed. Considering to special case of scenario-1, the optimum results were obtained from Scenario-3 in both equivalence class numbers and information loss ratios. In addition, the distribution of outlying class within all scenarios was confirmed that the scenario-3 gives acceptable values compared with the distributions of both equivalence class and information loss ratio. Figure 10 shows the distribution of outlying classes within all scenarios.

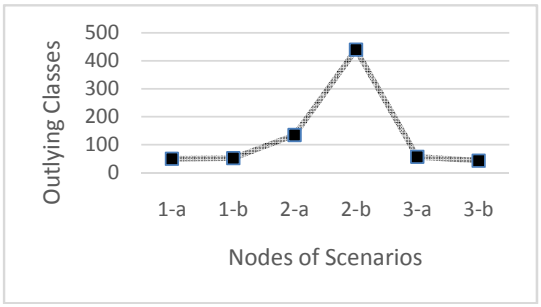


Figure 10: Outlying classes distribution of Scenarios

7. CONCLUSION

In this study, several privacy-preserving approaches were summarized and evaluated within the framework of health records. Our research consists of three main contributions in the three scenarios. In each role-based health data distribution scenario, only the required partition of personal health data is released for a specific organization by using database partitioning methods. Our proposed frameworks are practical to use different anonymization techniques with different metrics. From the experimental results, comparing the scenarios, the structure of scenario-3 is the most useful with respect to the data utility. Considering these architectures, many favorable structures will be generated for any field like commerce or banking. The future direction of the research is to generate an online query system for health records which have the architecture of scenario-3.

REFERENCES

Aggarwal C. C. 2005. On k-anonymity and the curse of dimensionality. In Proc. of the 31st Very Large Data Bases (VLDB), pages 901–909,Trondheim, Norway.

Aggarwal C.C., Yu P.S. 2008. Privacy-Preserving Data Mining: Model and Algorithms. Springer, Berlin.

Agrawal R., Srikant R. 2000. Privacy-Preserving Data Mining. ACM SIGMOD Conference

Bayardo R. J. and Agrawal R. 2005. Data privacy through optimal k-anonymization. In Proceedings of the 21st International Conference on Data Engineering (ICDE'05), pages 217 – 228. IEEE Computer Society.

Clifton C., Kantarcioglou M., Lin X., Zhu M. 2002. Tools for privacy-preserving distributed data mining. ACM SIGKDD Explorations, 4(2).

Dwork C., Nissim K. 2004. Privacy-Preserving Data Mining on Vertically Partitioned Databases, CRYPTO.

Dwork C. 2006. Differential privacy. In Proc. of the 33rd International Colloquium on Automata, Languages and Programming (ICALP), pages 1–12, Venice, Italy.

Fung B. C. M., Wang K., Chen R., and Yu P. S. 2010. Privacy-preserving data publishing: A survey on recent developments. ACM Computing Surveys, 42(4).

HIPAA. HIPAA. Health insurance portability and accountability act, 2004. http://www.hhs.gov/ocr/hipaa/.

Huang Z., Du W., Chen B. 2005. Deriving Private Information from Randomized Data. ACM SIGMOD Conference.

Jiang W., Clifton C. 2005. Privacy-preserving distributed k-Anonymity. Proceedings of the IFIP 11.3 Working Conference on Data and Applications Security.

Kohlmayer F., Prasser F., Eckert C., Kemper A. and Kuhn K. A. 2012. Flash: Efficent, Stable and Optimal K-Anonmity. ASE/IEEE International Conference on Social Computing, Amsterdam, Netherlands.

LeFevre K., DeWitt D., Ramakrishnan R. 2005. Incognito: Full Domain K-Anonymity. ACM SIGMOD Conference.

LeFevre K., DeWitt D., Ramakrishnan R. 2006. Workload Aware Anonymization. KDD Conference.

Lindell Y., Pinkas B. 2000. Privacy-Preserving Data Mining. CRYPTO.

Li N., Li T., and Venkatasubramanian S. 2007. t-closeness: Privacy beyond k-anonymity and l-diversity. In Proc. of the 21st IEEE International Conference on Data Engineering (ICDE), Istanbul, Turkey.

Machanavajjhala A., Gehrke J., Kifer D. 2006. -diversity: Privacy beyond k-anonymity. IEEE ICDE Conference.

Meyerson A., Williams R. 2004. On the complexity of optimal k-anonymity. ACM PODS Conference.

Nergiz M. E., Atzori M., and Clifton C. W. 2007. Hiding the presence of individuals from shared databases. In Proc. of ACM International Conference on Management of Data (SIGMOD), pages 665–676, Vancouver, Canada.

Newman D. J., Hettich S., Blake C. L., and Merz C. J. 1998. UCI repository of machine learning databases, http://archive.ics.uci.edu/ml/datasets/Adult

Pinkas B. 2002. Cryptographic Techniques for Privacy-Preserving Data Mining. ACM SIGKDD Explorations, 4(2).

Samarati P., Sweeney L. 1998. Protecting Privacy when Disclosing Information: k-Anonymity and its Enforcement Through Generalization and Suppression. IEEE Symp. on Security and Privacy.

Sweeney L. 2002. K-anonimity: A Model for Protecting Privacy. Int. J. Uncertainty, Fuzziness, Knowl-Based Syst. 10, 557-570.

Wang K., Fung B. C. M., Dong G. 2005. Integrating Private Databases for Data Analysis. Lecture Notes in Computer Science, 3495.

Zhong S., Yang Z., Wright R. 2005. Privacy-enhancing k-anonymization of customer data, In Proceedings of the ACMSIGMOD-SIGACT-SIGART Principles of Database Systems, Baltimore, MD.