Privacy Preserving Unstructured Data Publishing (PPUDP) Approach for Big Data

Ramya Shree A. N.
Research Scholar
CSE Department
RNSIT, Bengaluru-98

Kiran P.
Associate Professor
CSE Department
RNSIT, Bengaluru-98

ABSTRACT

The current trend related to Computer Industry is Big Data Analytics (BDA). The key attributes of Big Data are Volume, Variety, and Velocity. Big Data Analytics mainly focuses on Data collection from heterogeneous sources, Knowledge extraction from collected data and Data storage. In all these process of knowledge discovery from Big Data, Privacy of Big Data is very crucial. As it comprises of different types of data like structured data, unstructured data and semi structured data, a variety of techniques available for preserving privacy for structured and *semi structured data. The major issue with unstructured data is data publishing i.e. lack of preserving privacy because it may contain heterogeneous data like audio, video and text. The Big Data analytics normally carried out by third party, the data provider can be able to classify data as secured or not secured prior to data publishing for Data Analytics and it involves classification of huge volumes of data. The issue can be addressed using Privacy Preserving Unstructured Data Publishing for Big Data.

General Terms

PPDP, PPDM, MNB, KNN, NLP

Keywords

PPUDM, SDL

1. INTRODUCTION

The majority of Big Data is of type Unstructured namely audio, video, text etc. The unstructured data like text are difficult to analyze and contains knowledge that can be extracted by text mining techniques resulting in privacy threat. To avoid the threats text documents has to be transformed into a form to preserve privacy before publishing to Big Data Analytics process. The techniques used to preserve data privacy prior to data analysis called privacy preserving data publishing(PPDP).In order to preserve the privacy of the unstructured text data the first step is to apply pre processing techniques on text data and second is to categorize text documents into secured or not secured based on content. It requires domain knowledge to complete extent to determine secure features from text data.[1].The “unstructured” intends free form distribution of potentially sensitive data in text files with controls possibly become ineffective outside of specific systems or applications. Once the text file is shared, there is a possibility of high security risks. One such possible risk is untrusted public networks. Example customer or project details information in a text document sent by email from a public network other than specific organization network. Confidential customer references or reports downloaded to unrestricted internal servers and who has access for the server etc. This paper remaining sections organized like Section 2 illustrates about Architecture of PPUDP. Section 3 describes Unstructured Text Documents Preprocessing Method and example. Section 4 Explain about Text Documents Classification and Prediction. Section 5 presents the experimental results. Finally, Section 6 concludes this paper.

2. ARCHITECTURE OF PPUDP

The major concern related to Big Data Privacy is disclosure of sensitive data in the process of Knowledge Discovery from Big Data. The data collected from different heterogeneous sources are of high volume and it required prior processing to preserve data privacy before publishing for analysis. The three major roles involved in Knowledge Discovery Process from Big Data are- Data Provider, Data Collector and Data Analyzer. The Data Provider supply data to Data Collector/Data Analyzer through public network and Data Collector collects data from different data providers and supply to Data Analyzer for Analysis. Data Analyzer uses Data Mining Techniques to mine knowledge from data provided by Data Collector.

Data Analyzer uses PPDM methods to secure mining results. The Provider / Collector can control the sensitivity of the data when they provide data to Data Analyzer for data analysis. The provider should be able to explore his/her very private data, i.e. the data containing information that he/she does not want anyone else to know and inaccessible to the data collector or data analyzer. On the other hand, if the provider has to provide some data to the data collector, he/she wants to hide their sensitive information as much as possible.[2].In order to identify which are sensitive/not sensitive text documents at Data Provider level and Data Collector level we use our proposed technique. The privacy requirement identified at initial provider level before send/share data on public network such that individual privacy is protected. The traditional Privacy Preserving methods mainly operate on structured and semi structured data. [3][4].The proposed model used to preserve privacy of unstructured text data at Data Provider level and Data Collector level prior to Data Publishing.

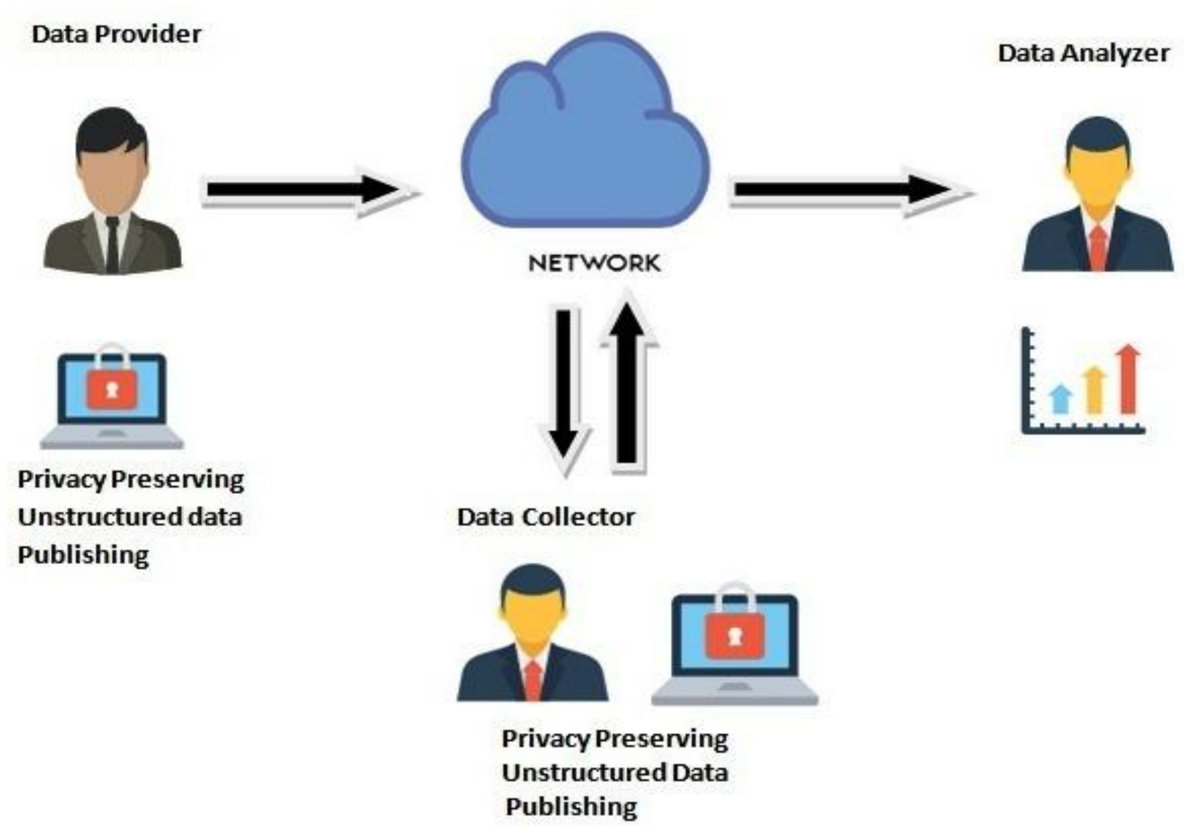
Data Provider
NETWORK
Data Analyzer
Privacy Preserving Unstructured data Publishing
Data Collector
Privacy Preserving Unstructured Data Publishing

Fig: 1 Architecture of Privacy

Preserving Unstructured Data Publishing at Data Provider/Data Collector

The key aspect of PPDP is to develop methods such that the sensitivity of data is not released. In this area various techniques are available and they tend to modify the original data in order to retain the sensitivity of data. The PPDP has been classified as following categories [7].

- Data Swapping - In this approach attributes of records are swapped with values such that it maintains the statistical inference in order to preserve privacy.
- Cryptography based PPDP - In this approach the Data is distributed in multiple sites and in order to mine the data must be securely retrieved from multiple sources. It can be treated as an efficient approach.
- Anonymization Approach - In this approach to preserve sensitivity modify the contents of the record owners before publishing the data. Popular techniques are k-anonymity, l-diversity-closeness.
- Randomization-In this approach noise is added to original attributes to mask attributes from disclosure. Randomization is done either by adding or multiplying noise.

3. TEXT DATA PRE PROCESSING

The unstructured text data has be converted into appropriate format before applying data mining techniques. The text documents pre-processing or Dimensionality Reduction (DR) allows an efficient data manipulation and representation. DR techniques are classified into Feature Extraction and Feature Selection approaches. The documents in text classification are represented by a great amount of features and most of them could be irrelevant or noisy. In Feature Extraction basic steps involved are: 1) Collect raw text data from different sources. 2) Perform data preprocessing on collected data it involves data cleaning and noise removal. The data further preprocessed to split sentences into words called tokens, text data contains stop words like “the”, “a”, “and” etc.Frequently occurring punctuations, headers and footers etc.The insignificant words need to be removed and represent words to stemmed words (root meaning) like healing to heal. The Natural Language Processing (NLP) is a technique which provides major operations as specified to perform text data preprocessing. [5][6][8].

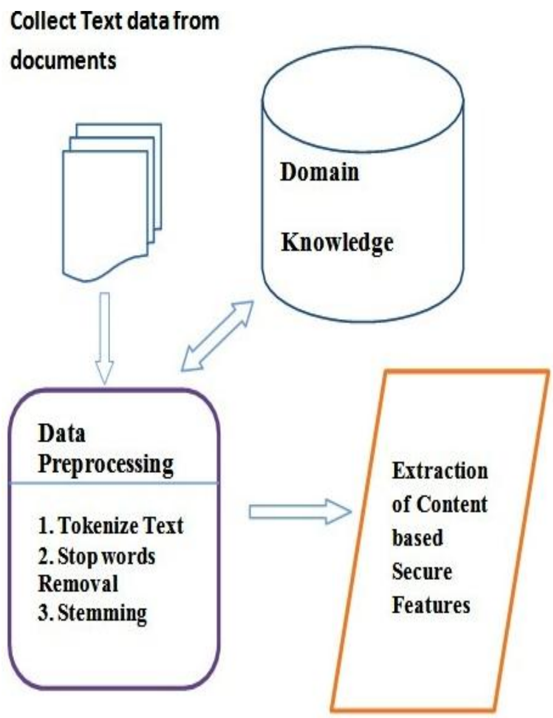

Fig: 2 The process of secure feature extraction from unstructured text documents

Original Raw Text Data Sample

Past medical history Nonalcoholic steatohepatitis Undergoing liver transplant workup. HAV, HCV reactive. HBV negative. Smooth muscle antibody positive. Large volume ascites with multiple paracenteses from 1/25 through 0/4. Liver biopsied on 6/14 with portal mononuclear inflammation, micro/macro vesicular steatosis, focal sinusoidal fibrosis.

'antibody', 'ascites', 'biop', 'fibrosis', 'focal', 'hbv', 'hcv', 'history’, ‘inflammation', 'large', 'liver', 'macro', 'medical', 'micro’, ‘mononuclear', 'multiple', 'muscle',' negative', 'nonalcoholic', 'paracenteses', 'past', 'plant', 'portal', 'positive', 'reactive', 'sied', 'sinusoidal', 'smooth', 'steatohepatitis', 'steatosis', 'through', 'trans', 'undergoing', 'vesicular', 'volume', 'workup'

Sample Result after Data Preprocessing

The Privacy Preserving methods focus on structured data but unstructured text in documents may also contains secure features that have to be identified and protected. If it not considered it may leads to privacy violation or information leak. Reserving Secure features are specific to domain and it requires full extent Domain Knowledge. Consider the example of medical domain where patient report contains disease details and patient personal information, when report shared or stored without considering the disease type and its associated description there is a possibility of personal privacy breach. In this case secure features are related to disease, disease type and patient who are having it. To avoid patient personalized privacy breach the patient report must be classified into secured or not secured prior to share or publish. To protect personalized privacy the patient report has to be associated with SDL (Sensitivity Disclosure Label).If SDL equals to 1 the document privacy must be preserved i.e. secured document otherwise SDL equals to 0 then the document is not secured [9][10].

4. CLASSIFICATION AND PREDICTION FROM UNSTRUCTURED TEXT

The text classification is a technique to assign predefined labels according to the content. Automatic text classification has many practical applications for example indexing for document retrieval, categorization of news articles, automatic metadata extraction, automatic identification of users reading interests, automatic e-mail sorting, and in general any application which requires document organization or selective and adaptive document dispatching. The text classification is the task of assigning a Boolean value to each pair <ti,cj > ∈ T

X C, where T is a domain of texts and C = {c1,c2,….ck} is a set of predefined categories. The Sensitivity Disclosure Label 1 is assigned to <ti,cj > indicates that the text document ti belongs to the category cj, and Sensitivity Disclosure Label 0 means that ti does not belong to the category cj. The task of identifying the category of each text is generally executed through a classifier that can be defined as a function ϕ: T X C →{1, 0} that approximates an unknown target.

The text classification used to extract secure feature from each document, and use the feature vectors as input to model secure classifier. Using set of predetermined words as features, word occurrence counts as feature values, a secure classifier is constructed for a set of documents. The documents, which belong to a SDL 1 are positive documents (True) and the remaining documents with SDL 0 are negative ones (False). The classifier considers the highest likelihood, predicts whether or not that SDL 1 is assigned to a new document based on the words in it, and the occurrence counts. These text categorization models that utilize words and their frequency as the features for training a classifier, is called vector space model. In this paper we explored the vector space model to build a classifier on the applied documents. To perform secure document classification we used Multinomial Naive Bayes and K-Nearest Neighbor classifiers [12][13][14].

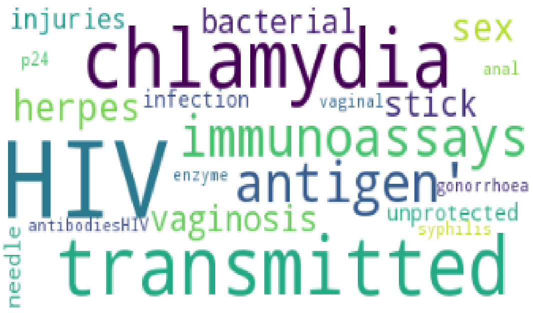

Fig. 3 Word cloud representation of Text Data

The Multinomial Naive Bayes (MNB) classifier is a probabilistic classifier based on Baye‟s Theorem with conditional independence assumptions. The MNB is an independent feature model. The independence assumptions of features make the features order is irrelevant and presence of a feature does not affect any another features in classification. It makes the computation of Bayesian classification approach more efficient, but it also limits its applicability. Depending on the precise nature of the probability model, the Multinomial Naive Bayes classifier can be trained efficiently by moderately little amount of training data to approximate the parameters necessary for classification.MNB specially used for text documents analysis. We used the MNB technique to perform secure classification [15].

The K-Nearest Neighbor (KNN) classifier is used to test the degree of similarity between preprocessed text documents and data, thereby determining the class of test documents. This method categorizes objects based on closest feature space in the training set. The training set is mapped into n dimensional feature space. The feature space is partitioned into regions based on the class of the training set. A point in the feature space is assigned to a particular class if it is the most frequent class among the k nearest training data.

Cosine similarity measure is used in computing the distance between the vectors. The key element of method is the usage of similarity measure for identifying neighbors of a particular document. Cosine similarity is dot product between vectors of two documents to find angle i.e. cosine of the angle to obtain similarity. The main consideration here is angle between two vectors. If the vectors are parallel to each other then documents are similar. If the vectors are orthogonal then documents are independent of each other [16][17][18].

5. RESULTS

The classification results obtained after proposed technique applied on i2b2 medical dataset. i2b2 (Informatics for Integrating Biology & the Bedside). The dataset contains patients discharge summary of type text documents. The proposed model developed using Python. The model predicts new medical discharge summary text document belongs to secured or not secured class i.e. sensitive or non sensitive.

The KNN classifier able to perform well when compared to Multinomial Naive Bayes classifier. The classifier model evaluation done by using Precision, Recall, f1-score and Accuracy measures. The evaluation measures calculated by using Confusion Matrix. It is summary of prediction results on a classification problem. The number of truthful and wrong predictions are summarized with count values and broken down by each class. The Precision, Recall, f1-score and accuracy measures calculated using equations[8]:

Precision = TP / (TP+FP) (1)

Recall = TP/ (TP+FN) (2)

f1-score = (2*Recall*Precision)/ (Recall + Precision) (3)

Accuracy= TP+TN /(TP+TN+FP+FN) (4)

Table 1. Comparison of MNB based on Classifier Model Evaluation

	Class	**Precision**	**Recall**	**f1-score**	**Accuracy**
MNB					
	SDL Positive	1.00	0.50	0.67	0.837
	SDL Negative	0.81	1.00	0.89	

Table 2. Comparison of KNN based on Classifier Model Evaluation

	Class	**Precision**	**Recall**	**f1 score**	**Accuracy**
KNN					
	SDL Positive	1.00	0.62	0.76	0.886
	SDL Negative	0.83	1.00	0.91	

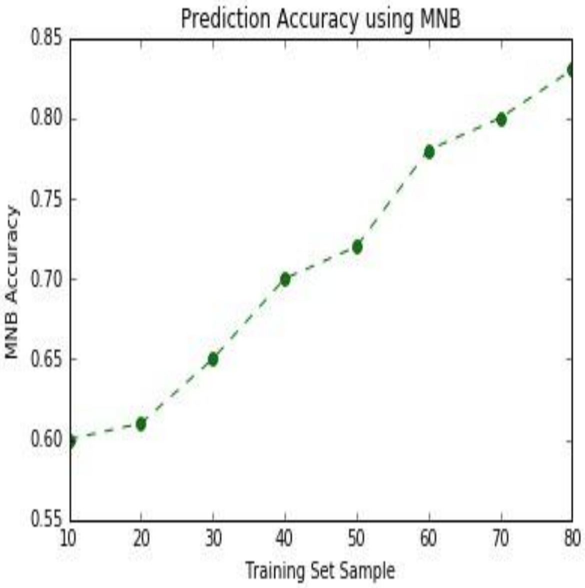

Fig. 4 Prediction Accuracy - MNB

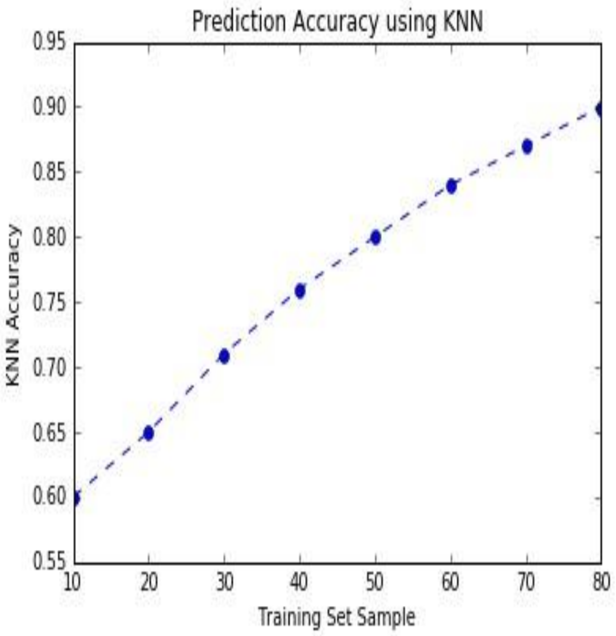

Fig. 5 Prediction A0ccuracy – KNN

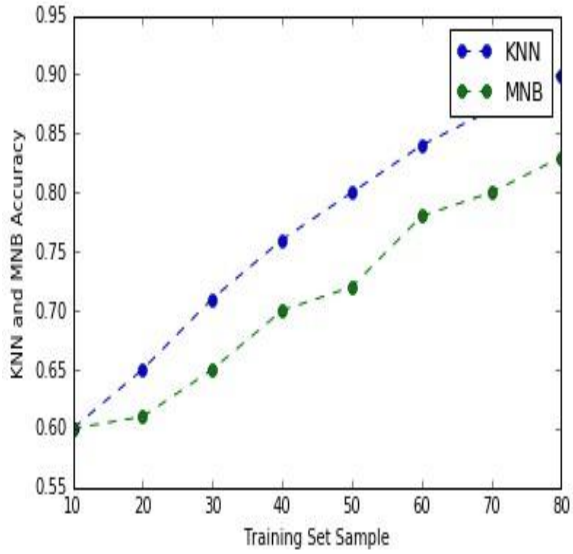

Fig.6 The KNN and MNB Classifier model Performance based on Accuracy measure

6. CONCLUSION

The Information Security is a vital issue in Big Data Applications. Preserving Individual/Personal privacy is a key requirement. Many standard techniques are available in Privacy Preserving Data Mining but it mainly concentrate on structured data. The different levels of users are involved in process of knowledge extraction and decision making at organizations. The major issue is heterogeneous data collected from different providers are integrated and used in process of knowledge extraction and decision making w.r.t to Big Data involves huge volume and variety of data. The unstructured data with data provider without privacy preserving concern shared to users (data collector, data analyzer) in the knowledge discovery may leads to security breach. The unstructured data with data collector without privacy preserving concern shared to data analyzer in the knowledge discovery may leads to security breach. To overcome in this paper an approach is proposed to privacy preserving data publishing for unstructured text documents (PPUDP) at data provider level and data collector level such that he/she can able to classify automatically secured or not secured document based on their domain privacy requirements. The data collector play major role and involves huge amount of data collection for Big data Analysis and it requires automatic classification of documents into secured not secured based on the application domain. Our approach can be used at Data Collector level with huge volumes of data and at Data Provider level to identify his/her documents to be published secured or not. The PPDP Techniques are used further once data classified into secured or not secured w.r.t to Knowledge Discovery Domain Requirements.

7. REFERENCES

[1] F. S. Gharehchopoh and Z. A. Khalifelu (2010) Analysis and Evaluation of Unstructured Data, IEEE Access.

[2] Lei Xu, Chunxiao Jiang,Jian Wang, Jian Yuan, & Yong Ren (2014) Information Security in Big Data: Privacy and Data Mining. V2, IEEE Open Access Journal.

[3] C. C. Aggarwal and S. Y. Philip (2012) A General Survey of Privacy-Preserving Data Mining Models and Algorithms. New York, NY, USA: Springer-Verlag.

[4] Ricardo Mendes, Joao P Vilela (2017) PPDM methods, metrics and applications. IEEE Access, Volume 5.

[5] F. Ronen and S. James (2006) The Text Mining Handbook: Advanced Approaches in Analyzing Data, vol. 1, Cambridge University Press.

[6] Manning C, Schutze H (1999) Foundations of Statistical Natural Language Processing, MIT Press, Cambridge, MA.

[7] Kiran P (2012) A Survey on Methods, Attacks and Metric for Privacy Preserving Data Publishing. International Journal of Computer Applications (0975-8887) Volume 53–No.18.

[8] Han, M. Kamber, and J. Pei (2006) Data Mining: Concepts and Techniques.San Mateo, CA, USA: Morgan Kaufmann Publishers.

[9] M. B. Malik, M. A. Ghazi, and R. Ali (2012) Privacy preserving data mining techniques: Current scenario and future prospects," in Proc. 3rd Int. Conf.Comput. Commun. Technol. (ICCCT), pp. 26-32.

[10] F.Popowich(2005) Using Text Mining and Natural Language Processing for Health Care Claims processing , SIGKDD Explorations, vol. 7, Issue 1, pp. 59-66.

[11] Gal, Tamas S, Z Chen, and A Gangopadhyay ((2008) A privacy protection model for patient data with multiple sensitive attributes. IGI Global 28–44.

[12] Dash .S, Vijayakumar K, Panigrahi B, Das S (2017) Artificial Intelligence and Evolutionary Computations in Engineering Systems. Advances in Intelligent Systems and Computing, vol 517. Springer, Singapore.

[13] Privacy-Preserving Data Mining: Methods, Metrics, and Applications ,IEEE Access, Volume 5,2017

[14] Privacy Preserving Data Mining Algorithms by Data Distortion,Wu Xiao-dan , Yue Dian-min ,Liu Feng-li , Wang Yun-feng ,Chu Chao-Hsien, ,International Conference on Management Science and Engineering, 04

September 2007,IEEE Access,Volume5

[15] Naive Bayes Classifier, Lecture Notes in Computer Science, 2005, Vol 3584

[16] Privacy preserving data mining - „A state of the art‟ Mamta Narwaria , Suchita Arya 3rd International Conference on Computing for Sustainable Global Development, 2017 [17]. S. Matwin, "Privacy-preserving data mining techniques: Survey and challenges", in Discrimination and Privacy in the Information Society, Berlin, pp. 209-221, 2013.

[17] S.-W. Chen et al.,Confidentiality protection of digital health records in cloud computing,‟‟‟ J. Med. Syst., vol. 40, no. 5, p. 124, 2016. [95]

[18] J. Vincent, W. Pan, and G. Coatrieux, „„„Privacy protection and security in ehealth cloud platform for medical image sharing,‟‟‟ in Proc. IEEE 2nd Int. Conf. Adv. Technol. Signal Image Process. (ATSIP), Mar. 2016, pp. 93–96. [96]

[19] M. S. Kiraz, Z. A. Genç, and S. Kardas, „„„Security and efficiency analysis of the Hamming distance computation protocol based on oblivious transfer,‟‟‟ Secure. Commun. Netw., vol. 8, no. 18, pp. 4123–4135, 2015. [97]

[20] J. Zhou, Z. Cao, X. Dong, and X. Lin, „„„PPDM: A privacy-preserving protocol for cloud-assisted e-healthcare systems,‟‟‟ IEEE J. Sel. Topics Signal Process., vol. 9, no. 7, pp. 1332–1344, Oct. 2015. [98]

[21] P. J. McLaren et al., „„„Privacy-preserving genomic testing in the clinic: A model using HIV treatment,‟‟‟ Genet. Med., vol. 18, no. 8, pp. 814–822, 2016. [99]