A Review Of Synthetic Data Generation Methods For Privacy Preserving Data Publishing

Surendra .H, Dr. Mohan .H .S

Abstract: Due to the technological advancement, enormous micro data containing detailed individual information is being collected by both public and private organizations. The demand for releasing this data to public for social and economic welfare is growing. Also the organizations holding the data are under pressure to publish the data for proving their transparency. Since this micro data contains sensitive information about individuals, the raw data needs to be sanitized to preserve privacy of the individuals before releasing it to the public. There are different types of data sanitization methods and many techniques are being proposed for Privacy Preserving Data Publishing (PPDP) of micro data. Synthetic Data Generation is an alternative to data masking techniques for preserving privacy. In this paper different fully and partially synthetic data generation techniques are reviewed and key research gaps are identified which needs to be focused in the future research.

Index Terms: Disclosure Control, Data Masking, Inference Control, Privacy Preserving Data Publishing (PPDP), Privacy Preserving Data Mining (PPDM), Synthetic Data Generation.

1 INTRODUCTION

Recent advancements in internet and computing technologies have made it feasible to collect, store and process large amount of micro data by the organizations. This micro data contains detailed information of individual entities which may also contain their private and sensitive information. Due to the collaborative nature of business in many enterprises, this data needs to be shared among multiple third parties. Also to support research and innovation, organizations need to share this data to the public. Sharing the raw micro data to any third party pose risk of disclosing private information of the individual entities which can be misused by adversaries. So the raw data needs to be sanitized to protect privacy and reduce the disclosure risk. The raw data is sanitized using disclosure control methods which can be broadly categorized as Perturbative and Non-Perturbative methods. In Perturbative methods, the data raw is added with noise such a way that sensitive information is masked retaining key statistical information. In the Non-Perturbative methods, the data is generalized such that no individual entity can be particularly identified. Synthetic Data Generation is another technique where the private and sensitive data in the original data is replaced with the synthetic data. Also instead of releasing the processed original data, complete data to be released can be fully generated synthetically. Synthetic Data Generation has taken focus in recent years not only for its usage in privacy preserving data publishing but also for its capability to support validation of new algorithms and applications which needs data which is not available or not accessible due to privacy concerns. In this paper different data sanitization methods based on synthetic data generation for privacy preserving data publishing are reviewed and the key findings of the review with respect to its practical applications are discussed. Also the future scope of research in this field is presented.

- *Surendra .H is currently pursuing PhD degree program in privacy preserving data publishing and mining in SJB Institute of Technology, Bangalore, India, E-mail: surendra.h@gmail.com*
- *Dr. Mohan .H .S is currently working as professor and head of Department of Information Science at SJB Institute of Technology, Bangalore, India, E-mail: mohan_kit@yahoo.com*

The organization of the paper is as follows. Section II briefs different synthetic data generation types. Section III provides review of different synthetic data generation methods used for preserving privacy in micro data. Section IV discusses about the key findings of the study and list out the important characteristics that a synthetic data generation method shall posses for protecting privacy in big data. The paper is concluded in Section V.

2 TYPES OF SYNTHETIC DATA

Synthetic data generation is an alternative data sanitization method to data masking for preserving privacy in published data. The data is randomly generated with constraints to hide sensitive private information and retain certain statistical information or relationships between attributes in the original data. The published synthetic data are broadly classified into three categories [1];

2.1 Fully Synthetic Data

This data is completely synthetic and doesn't contain original data. The fully synthetic data generators identify the density function of attributes in the original data and estimate the parameters of these density functions. Then for each attribute, privacy protected series are generated by randomly picking up the values from the estimated density functions. If only few attributes of the original data are selected for replacing with synthetic data then the protected series of these attributes are mapped with the other attributes of the original data to rank the protected series and the original series in same order. Multiple imputation and bootstrap methods are few classical techniques used to generate fully synthetic data. Since the released data is completely artificially generated and doesn't contain original data, this technique has strong privacy protection but the truthfulness of the data is lost.

2.2 Partially Synthetic Data

In contrast to the fully synthetic data, the method used to generate partially synthetic data replaces only values of the selected sensitive attribute with synthetic values. The original values are replaces only if it posses high risk of disclosure. Masking the original values with synthetic values prevents re-identification thus preserving privacy in the published data. Multiple imputation and model based techniques have been used to find the synthetic values for

the selected attribute to avoid disclosure. These techniques are also useful for imputing missing values in the original data. Disclosure risk is higher in partially synthetic data compared to fully synthetic data as it contains original data along with imputed synthetic data.

2.3 Hybrid Synthetic Data

The hybrid synthetic data is generated using both original and synthetic data. For each record of original data a nearest record in the synthetic data is chosen and both are combined to form hybrid data. The hybrid synthetic data holds the advantages of both fully and partially synthetic data. Hence it will provide good privacy preservation with high utility compared to fully synthetic and partially synthetic data but at the cost of more memory and processing time.

3 Survey of Synthetic Data Generation Methods

The purpose of this survey is to study different synthetic data generation methods and identify research gaps. The survey is not an exhaustive study but includes recent developments in generation of synthetic data for privacy preserving data publishing. The comparative study of these techniques is given in Table 1.

TABLE 1

Comparative Study of Synthetic Data Generation Methods

Author, Reference	Year	Data Generation Type	Proposed Method	Limitations
Pengyue J. Lin, Behrokh Samadi and Daniel R. Jeske [2]	2006	Fully Synthetic	• Semantic Graph based synthetic data generation is proposed for testing Information Discovery System. • The semantic graph is developed using three different kinds of rules which are Independent Rules, Intra-record rules and Inter-record rules. • Independent Rules govern the values of the attributes independent of other attribute or records. • Intra-record rules govern the values of attributes in relation to the other attribute values of the same record. • Inter-record rules govern the values of attributes in relation to the attribute values of other records in the dataset. • These rules/relations are represented as graph structure and synthetic data is generated satisfying these rules.	• The data is generated by the rules defined by the user and not learnt from any real data. • A large number of rules need to be defined even for small attributes and range of values. Suffers from dimensionality curse. • User should know the list of attributes and all possible values of those attributes prior to define the rules. • As data is generated only by user defined rules and doesn't represent any real data, the truthfulness of the data is completely lost. • The utility of this data is limited to testing Information Discovery Systems.
Daniel R. Jeske and Behrokh Samadi [3]	2006	Fully Synthetic	• Extended the method proposed in [1] to include scenario insertion and re-sampling techniques to generate close to real data set for testing data mining tools.	• Inherits the limitations of [2]
Jilles Vreeken, Matthijs van Leeuwen and Arno Siebes [4]	2007	Fully Synthetic	• Proposed Minimal Description Length (MDL) based KRIMP algorithm to generate privacy preserved data. • The algorithm uses coding table to encode the transaction database in compact form. • The support values of the itemsets in the coded database are Laplace corrected. The generator picks itemset with high frequency randomly from the code table and generated transactions. • The represented database is very compact.	• The algorithm is very time intensive as it has to code each itemsets. • Not suitable for streaming data as it is difficult to determine high frequency itemsets used to encoding shorter length code due to concept drift. • The results are approximated even for original transactions which are not sensitive.
Josh Eno and Craig W. Thompson [5]	2008	Fully Synthetic	• Proposed a method to reverse generate the synthetic data from decision tree model. • Used Synthetic Data Definition Language (SDDL), an XML based language to represent the data model. • The Predictive Model Markup Language (PMML) representation of decision tree is used to generate synthetic data in SDDL format.	• The decision tree used can be built using real data. So while reverse generating the synthetic data from the decision tree, privacy preservation of any individual is not considered. • The decision tree is built with a particular attribute at class variable. The data generation process cannot be personalized to the user needs i.e. user may want some other attribute as class variable instead of the one used to build the decision tree.
Isaac Cano and Vicenc Torra [6]	2009	Partially Synthetic	• Proposed Fuzzy c-Regression Models (FCRM) based on Fuzzy c-Means Clustering algorithm.	• Designed for numeric and continuous variables and doesn't work for categorical variables.

			• The attributes are divided into independent and dependent variables. Then the records are clustered to a user defined number of clusters c. For each record, the maximal membership to any cluster centroid is found. Then using regression technique the values of the dependent variables are estimated and replaced with the original values while generating synthetic data. • The values of the independent variables are unchanged. The level of information loss or the disclosure risk can be controlled by the number of clusters. Higher the cluster count, higher the discloser risk. The evaluation on the information loss with crisp and fuzzy partition is presented in [6]	• If the level of disclosure risk is changed then the whole algorithm needs to be restarted with new c value (number of cluster) which reduces the overall time efficiency of the algorithm.
Jerome P. Reiter and Jorg Drechsler [7]	2010	Fully Synthetic Partially Synthetic	• Proposed two-stage process to generate synthetic data. • In the first stage samples are drawn from the original dataset using stratified sampling. Then some attribute values are imputed using multiply imputation technique. In the second stage the remaining attribute values are imputed with the use of attribute values imputed in the first stage. • This method is also helpful in imputing missing values. Can be used for generating both fully synthetic and partially synthetic data.	• Considers samples from the original data for modeling which will reduce the accuracy of the model. • Two stage of imputation decreases the time efficiency of the system.
Jorg Drechsler [8]	2010	Fully Synthetic Partially Synthetic	• Proposed Support Vector Machine (SVM) based synthetic data generation. • The SVM is trained with original data to create a classifier model. The trained model is then used to generate synthetic data. • It is an initial investigation to check the feasibility to use SVM for synthetic data generation.	• The disclosure risk is very high if the classification accuracy of the SVM is more. • Applicable for numeric and continuous data and not tested for categorical data.
Gregory Caiola and Jerome P. Reiter [9]	2010	Partially Synthetic	• Proposed generating partially synthetic data using Random Forest • The Random Forest ensemble classifier is used to model the original data. Then based on the requirement, the classifier model is tuned to generate privacy preserved partial synthetic data.	• Designed for categorical variables and doesn't take care for numerical and continuous attributes. • The structure of the Random Forest depends on the class variable selected. If the user requires another attribute as class variable then the algorithm needs to re-run from scratch to generate the required synthetic data for a user.
Georgia Albuquerque, Thomas Lowe and Marcus Magnor [10]	2011	Fully Synthetic	• Proposed a framework to generate multivariate high dimensional fully synthetic data by sampling statistical distributions in a multidimensional space. • The user defines the number of dimensions required and structure of the multidimensional data through probability density functions. The data points are generated by sampling from these probability density functions.	• Doesn't use real data. Prune to bias introduced by the user while defining the probability distribution functions. • Processing time exponentially increases with the dimension of the data
Arvind Arasu, Raghav Kaushik and Jian Li [11]	2011	Fully Synthetic	• Proposed generation of synthetic data by specifying cardinality constraints. • User specifies the database schema and the cardinality constraints what the synthetic data generation shall obey while generating the data.	• Doesn't use real data. • Doesn't handle overlapping constraints.
Moritz Hardt, Katrina Ligett and Frank McSherry [12]	2012	Fully Synthetic	• Proposed Multiplicative Weights update rule with Exponential Mechanism (MWEM) which is an extension to Exponential Mechanism to generate differentially private synthetic data. • The Multiplicative Weights approach is used to scale up or down the weights of the records in the approximate data which contribute to the accuracy of the query result. The approximation is improved repeatedly to find the best approximate data release which gives result closed to the original data.	• The set of queries shall be pre-determined by the user and the differential data released is accurate only to these queries. • Since it has to find good approximation of the original data, it takes very longer to time to converge for a large dataset.

Noman Mahammed, Xiaoqian, Rui Chen, Benjamin C M Fung and Lucila Ohno-Machado [13]	2013	Fully Synthetic	• Proposed a differentially private data release method where the data is first generalized and then noise is added. • First, the predictor attributes are generalized and partitions into several equivalence classes. Then noise is added and publishes the noisy counts of the groups.	• Not feasible for large datasets • Finding optimum epsilon and specialization value is difficult.
Marden Pasinato, Carlos Eduardo Mello, Marie-Aude Augaure and Geraldo Zimbrao [14]	2013	Fully Synthetic	• Proposed a methodology to generate fully synthetic data to evaluate context aware recommendation systems. • Their methodology generate user, product and their association using some user defined probability density functions. • A penalization function alters the ratings given by the user if it differs too much from the user context. This way outliers are also treated.	• Doesn't use real data. Biased by the probability density functions that the user specifies. • The penalization function may take more time to converge to an optimum rating which increases the accuracy of the data and may be sensitive to outliers.
Yubin Park, Joydeep Ghosh and Mallikarjun Shankar [15]	2013	Fully Synthetic	• Proposed a fully synthetic data generation method using perturbation of data and Gibbs sampling. • In the first step, the algorithm estimates the approximate probability distribution of attribute values. Then the data is distorted using any perturbation technique like l-diversity or differential privacy. Finally, using Gibbs sampling technique, the required size of records is sampled from the perturbed data. • A parallel execution of the proposed algorithm is also presented which can be used in distributed computing environment to generate big data.	• Doesn't use real data. Biased by the probability density functions that the user specifies. • Sampling will further reduce the utility of the final released data as the data from it is sampling is fully synthetic. • The execution time increases exponentially with the increase of dimensionality of the data.
Jun Zhang, Graham Cormode, Cecilia M. Procopiuc, Divesh Srivastava and Xiaokui Xiao [16]	2014	Fully Synthetic Partially Synthetic	• Proposed a differentially private method for releasing high dimensional data using Bayesian Network. • The Bayesian Network is used to model the relationship between the attributes in the data. • The distribution of the data in original data is approximated using low dimensional marginal and then this data is added with noise to satisfy differential privacy. Finally, samples are drawn from the differentially private data for release.	• The Bayesian Network grows to very large graph for data with very large dimension like retails data where there will be millions of products. • Adding noise before sampling the data will reduce the utility of the sample data.
Wentian Lu, Gerome Miklau and Vani Gupta [17]	2014	Fully Synthetic	• Proposed a two phase differentially private data release mechanism. • The data owner creates a data model satisfying statistical characteristics required for serving given set of queries. Then the model is perturbed using differential privacy technique and released to the public. • The data user can then generate synthetic data to work on his queries.	• The data modeled is specific to the query workload provided by the user. If a new set of queries need to be answered or few queries in the existing workload need to be modified, then the model shall be re-generated satisfying these new queries. • Increased processing time for high dimensional data and large dataset as the perturbation technique used is a repetitive method to get the optimal set of records for publishing.
Yubin Park and Joydeep Ghosh [18]	2014	Fully Synthetic	• Proposed Perturbed Gibbs Sampler (PeGS) method to generate categorical synthetic data. • The data is disintegrated to create blocks of data, each block representing statistical characteristics of the original data. Then these statistical blocks are modified to add required level of noise. Then the synthetic data is generated by sampling from the modified statistical blocks.	• The data modeling used is not memory efficient as it stores each and every statistical information of data. • Perturbed Multiple Imputation will increase overall processing time for high dimensional data.
Haoran Li, Li Ziong, Lifan Zhang and Xiaoqian Jiang [19]	2014	Fully Synthetic	• Developed a Differentially Private Data Synthesizer named DPSynthesizer for Privacy Preserving Data Publishing. It provides easy to use web interface to synthesize data. • The Synthesizer models the data by first generating histogram for each attribute in the original dataset and then the dependency matrix is generated using Gaussian copula function on the original data. • The differentially private synthetic data is generated	• When new set of data need to be added to the original data then computation of dependency matrix will be time consuming. So updates are very expensive. • Efficiency drops considerably for high dimensional large dataset such as big data.

			by random sampling from the marginal histograms and the dependency matrix.	
Matthew J. Schneider and John M. A [20]	2015	Fully Synthetic	• Proposed Bayesian method based data modeling with zero-inflation. This method doesn't require suppressing the data in order to protect confidentiality.	• Applicable for numeric data and not demonstrated for categorical data.
Harichandan Roy, Murat Kantarcioglu and Latanya Sweeney [21]	2016	Fully Synthetic	• Proposed generation of differentially private synthetic human behavior data. • Contingency Table containing histogram of all the attributes are generated and then Laplace noise is added to each histogram to make them differentially private. • The privacy preservation is done by converting the sensitivity from presence or absence of individual to presence or absence of trip.	• Applicable to human trip data and not generic. • Doesn't support personalization of the generation of synthetic data. • Utility of the data decreases for high dimensional data as more noise is added while creating contingency table.
Vanessa Ayala-Rivera, A. Omar Portillo-Dominguez, Liam Murphy and Christina Thorpe [22]	2016	Fully Synthetic	• Developed a framework for generating synthetic micro dataset. • The user defines the rules and constraints needed to preserve the functional dependencies of any given business case. Also the list of attributes and their domain are specified by the user. • The algorithm iteratively generates the synthetic data from different type of generators satisfying the user defined rules and constraints of a given size. The generators can use different probability density functions to satisfy the functional dependencies of the given business case.	• User should be a domain expert to specify proper rules, constraints, attributes and their values which is not normally the case. • Doesn't use real data. Biased to the specification of the user. • The iterative process reduces the time efficiency for high dimensional, large dataset.
David F. Nettleton and Julian Salas [23]	2016	Fully Synthetic	• Proposed anonymized synthetic generation of social network data. • The user specifies the size of the social network with other rules and constraints such as maximum number of local neighbors for any user, information propagation, ontology etc. • The user also specifies the level of privacy requires by providing K value for k-anonymity and T value for t-closeness. • The data generator generates the data as per user specification and the data anonymizer synthesizes the generated data to satisfy the privacy parameters set by the user.	• Doesn't use real data. • Pre-determination of sensitive attributes limits the flexibility of usage of the data. The selection of sensitive attribute shall be given to the end user based on which the anonymization shall take place. • Addition or deletion of users in the social network is not taken care. This is required to study the dynamics of the social network.

4 KEY FINDINGS OF THE SURVEY

This survey includes prominent research work carried out for the purpose of generating synthetic data without the need for the real data. Since there is no real data used, the issue with privacy breach doesn't exist. But also the truthfulness of the data is also lost. We have considered only the techniques used in these work and not the purpose, but still most of the work considered are used for preserving privacy in the released data through synthetic data generation. Most of these algorithms are designed for generating fully synthetic data and generation of partially synthetic data is limited. This might be due to increased disclosure risk in release of partially synthetic data. Other key findings of the survey are given below.

4.1 Data Modeling Technique

Techniques that generate synthetic data without original data use set of rules and constraints defined the user. The data is modeled using set of rules and relationships. The user needs to have good understanding of the domain which the generated data represents. Techniques that use the original data for generating the fully or partially synthetic data learn the characteristics of the data and model's it using histogram or probability density functions. These methods shall be scaled to handle high dimensional large data like retail transactions that can have millions of products and customers. Also the data model shall aid development of privacy preserved data mining techniques such as association rule mining, classification, clustering etc using the functions of the model itself and not by learning from the synthetic data generated from the model. However, the generated synthetic data can be used to train a data mining algorithm if the model based algorithms are not satisfactory. The data structure used to store the model shall be memory and time efficient so that it can be used for real-time applications. Also the data structure shall be able to scale to distributed computing and storage environment for processing Big Data.

4.2 Personalized Privacy Preservation and Data Generation

While modeling the original data, all statistical and functional characteristics of the data are stored. During generation of the synthetic data from the model, the privacy preservation is taken care. As we can re-generated the original data from the model, most of time users need only

part of the data. It might be less number of attributes than the original data contains or constrained on the value of one or more attributes. User shall be allowed to personalize the generation of data, of course with preservation of privacy. The synthetic data generator preserves privacy pertaining to the list of attributes or constraints the user has specified. This will increase the utility of the generated data since information loss due to privacy preservation is less compared to the information loss that have occurred to privacy preservation of complete original data. As some algorithms support specification of size of the data generated, the data model shall provide simple method to scale up or scale down the size of the data generated. Gibbs, random and other sampling techniques are being used to sample the data from the model; it shall be improved to represent the original data as accurately as possible without compromising on the privacy.

4.3 Partially Synthetic Data Generation

From the survey it is evident that most of the techniques discussed are designed for generating fully synthetic data and there are very limited techniques which support partially synthetic data generation. Even though the disclosure risk is higher in partially synthetic data, the truthfulness of the data is high. Whereas fully synthetic data has strong resistance to disclosure risk but lacks truthfulness. Fully synthetic data are very useful in research but has limited scope for usage in commercial applications. From the point of commercial usage of the data, a more robust privacy preserved partially synthetic data generation technique shall be developed.

4.4 Modeling Data Stream

Modeling streaming data is more challenging than modeling static data as the distribution of the data is not know prior. In multivariate data, the relationship among the attributes may vary with time and these changes need to be captured in the model. The information loss due to privacy preservation while generating the data depends on the characteristics of the model at the point of generation. Special care must be taken to avoid disclosure in incremental release of the data. As most of the techniques studied are designed to model static data, the data modeling technique shall support incremental learning and update to the model with latest data records. The data generation shall take care of preserving privacy in incremental release of the data along with privacy preservation in the whole released data.

5 Conclusion

Synthetic data generation is one of the important technique for privacy preserving data publishing. As the published data doesn't represent any real entity, the disclosure of sensitive private data is eliminated. If the information available in the released synthetic data matches with any real entity participated in the original data then it is purely a co-incidence unless the data is partially synthetic. For release of partial synthetic data, stringent privacy protection techniques shall be used as part of it contains information from original data. Hybrid synthetic data contains both original and synthetic data but not in practice due to its computational complexity. The earlier techniques were designed around anonymization techniques for preserving privacy, recent are using differential privacy which is yet to mature for practical usage in commercial applications. Synthetic data generation has got more focus in the recent years as it helps in validating new techniques and technologies quickly without wanting to have original data. So there is a need for developing techniques to precisely model the data either by original data or by user defined specifications and allowing the end user to personalize the data generation as per his needs while ensuring privacy of the individual entities contributing to the data in any. In this paper we studied different synthetic data generation techniques proposed for publishing privacy preserved data and the future scope of the research in the field of synthetic data generation is discussed.

References

[1] Charu C. Aggarwal and Philip S. Yu, Privacy-Preserving Data Mining: Models and Algorithms, Springer series in Advances in Database Systems, vol. 34, 2008.

[2] P.J.Lin, B. Samadi, A. Cipolone, D.R. Jeske, S.Cox, D. Holt and Rui Xiao, "Development of Synthetic Data Set Generator for Building and Testing Information Discovery Systems", in Proc. of the Third Int. Conf. on Information Technology: New Generations, Apr 2006.

[3] Daniel R. Jeske, Pengyue J. Lin, Carlos Rendon, Rui Xiao, Behrokh Samadi, "Synthetic Data Generation Capabilities for Testing Data Mining Tools", IEEE Military Communications Conference, Oct 2006.

[4] Jilles Vreeken, Matthijs van Leeuwen and Arno Siebes, "Preserving Privacy through Data Generation", in Proc. of Seventh IEEE Int. Conf. on Data Miing, Oct 2007.

[5] Josh Eno and Craig W. Thompson, "Generating Synthetic Data to Match Data Mining Patterns", IEEE Internet Computing, vol. 12, Issue 3, May-June 2008, pp. 78-82.

[6] Issac Cano and Vicenc Torra, "Generation of synthetic data by means of fuzzy c-Regression", in Proc. of IEEE Int. Conf. on Fuzzy Systems, Aug 2009, pp. 1145-1150.

[7] Jerome P. Reiter and Jorg Drechsler, "Releasing Multiply-Imputed Synthetic Data Generated in Two Stages to Protect Confidentiality", Statistica Sinica, vol. 20, no. 1, Jan 2010, pp. 405-421.

[8] Jorg Drechsler, "Using Support Vector Machines for Generating Synthetic Datasets", in Proc. of Int. Conf. on Privacy in Statistical Databases, 2010, pp. 148-161.

[9] Gregory Caiola and Jerome P. Reiter, "Random Forests for Generating Partially Synthetic, Categorical Data", Trans. Data Privacy, Apr 2010, pp. 27-42.

[10] Georgia Albuquerque, Thomas Lowe and Marcus Magnor, "Synthetic Generation of High Dimensional Data", IEEE Trans. on Visualization and Computer Graphics, vol. 17, issue 12, Dec 2011, pp. 2317-2324.

[11] Aravind Arasu, Raghav Kaushik and Jian Li, "Data Generation using Declarative Constraints", in Proc. of ACM SIGMOD Int. Conf. on Management of Data, Jun 2011, pp. 685-696.

[12] Moritz Hardt, Katrina Ligett and Frank Mcsherry, "A Simple and Practical Algorithm for Differentially Private Data Release", in Proc. of Advances in Neural Information Processing Systems, 2012, pp. 2339-2347.

[13] Noman Mohammed, Xiaoqian Jiang, Rui Chen, Benjamin C M Fung and Lucila Ohno-Machado, "Privacy-preserving heterogeneous health data sharing", Journal of the American Medical Informatics Association, vol. 20, issue 3, May 2013, pp. 462-469.

[14] Marden Pasinato, Carlos Eduardo Mello, Marie-Aude Aufaure and Geraldo Zimbaro, "Generating Synthetic Data for Context-Aware Recommender Systems", BRICS Congress on Computational Intelligence & 11th Brazilian Congress on Computational Intelligence, Sep 2013, pp. 563-567.

[15] Yubin Park, Joydeep Ghosh and Mallikarjun Shankar, "Perturbed Gibbs Samplers for Generating Large-Scale Privacy-Safe Synthetic Health Data", in Proc. of IEEE Int. Conf. on Healthcare Informatics, Sept 2013, pp. 493-498.

[16] Jun Zhang, Graham Cormode, Cecilia M. Procopiuc, Divesh Srivastava and Xiaokui Xiao, "PrivBayes: private data release via Bayesian Networks", in Proc. of ACM SIGMOD Int. Conf. on Management of Data, Jun 2014, pp. 1423-1434.

[17] Wentian Lu, Gerome Miklau and Vani Gupta, "Generating private synthetic databases for untrusted system evaluation", in Proc. of IEEE Int. Conf. on Data Engineering, Mar-Apr 2014, pp. 652-663.

[18] Yubin Park and Joydeep Ghosh, "PeGS: Perturbed Gibbs Samplers that generate Privacy-Compliant Synthetic Data", Trans. on Data Privacy, vol. 7, issue 3, Dec 2014, pp. 253-282.

[19] Haoran Li, Li Xiong, Lifan Zhang and Xiaoqian Jiang, "DPSynthesizer: Differentially private data synthesizer for privacy preserving data sharing", in Proc. of the VLDB Endowment, vol 7, issue 13, Aug 2014, pp. 1677-1680.

[20] Matthew J. Schneider and John M. Abowd, "A New Method for Protecting Interrelated Time Series with Bayesian Prior Distributions and Synthetic Data", Journal of the Royal Statistical Society, Series A, 2015.

[21] Harichandan Roy, Murat Kantarcioglu and Latanya Sweeney, "Practical Differentially Private Modeling of Human Movement Data", in Proc. of Thirtyth conf. on Data and Applications Security and Privaty, vol. 9766, Jul 2016, pp. 170-178.

[22] Vanessa Ayala-Rivera, A. Omar Portillo-Dominquez, Liam Murphy and Christina Thorpe, "COCOA: A Synthetic Data Generator for Testing Anonymization Techniques", in Prod. of Int. Conf. on Privacy in Statistical Databases, 2016, pp. 163-177.

[23] David F. Nettleton and Julian Salas, "A data driven anonymization system for information rich online social network graphs", Int. Journal of Expert Systems with Applications, vol. 55, issue C, Aug 2016, pp. 87-105.