Differentially Private Data Release for Data Analytics - A Model Review

Peter N. MUTURI[1], Andrew M. KAHONGE[1], Christopher K. CHEPKEN[1]
[1]School of Computing and Informatics,
University of Nairobi,
Nairobi, Kenya.
pmuturi@mmu.ac.ke, andrew.mwaura@uonbi.ac.ke, chepken@uonbi.ac.ke

To leverage on the potential of data analytics, enabling private data release is needed. The challenge in achieving private data release has been balancing between privacy and analytical utility. Among the models that seek to solve the challenge, ε-differential privacy promises to achieve the balance by regulating the epsilon (ε) value. The choice of the appropriate epsilon value that achieves the balance has been a challenge, making the ε-differential privacy not practically applicable by many. A practical and heuristic method to estimate this privacy parameter needs formulation. The variable to estimate appropriate privacy parameter that is not provided in heuristic manner is the reidentification probability. Previous research has based that probability on released data sets and linkage data sets, with less focus on data analysts. This paper proposes a causal relationship model for estimating the reidentification probability, which adds the analyst's aspect to the model.
Keywords: *Privacy, Data Utility, Differential Privacy, Big Data, Private release, Anonymization*

1 Introduction

With the convergence and working together of smart devices, the Internet of Things, and Internet-based applications, massive data can be produced, collected, processed, and stored effectively. However, storing large volumes of data without making value from it is not helpful, and, indeed, it is a waste of computing resources [1]. The value of the data sets held is achieved through data analytics [2], which may be undertaken by the owners of these data sets (curators), or the curators may release the data sets, which are then used by third parties for various analytical purposes. One concern that needs to be addressed to actualize the data release for data analytics is how the private information contained in the data sets should be protected [3].

Without data collecting agencies (curators) that guarantee the protection of private information held in the data sets that they store, only a few individuals would willingly participate in any data collection exercise, and those who do participate may not provide very accurate data [4]. As observed by Cavoukian and Reed in [5], the challenge of safeguarding privacy threatens the willingness to release data and information. However, with a mechanism that provides the guarantee that the privacy of individuals in the data sets is in place, thereby creating anonymous data sets, then a framework that allows data curators to make such data available to third parties or even to the public for analytical purposes can be put in place. This is what is known as private data release. Analysts who interact with such private data can only learn about the population from which the data was obtained, but not about an individual whose data is in the dataset.

The process of making data private involves suppression, aggregation, noise addition, swapping, among other mechanisms [6]–[8]. This, in effect, affects the analytical value of the data, hence reducing the data's analytical utility. If ensuring privacy of the data sets was the only goal, this would be achieved trivially [9]. However, in pursuing high levels of privacy through various mechanisms, data sets may end up losing the analytical utility that is very core for data analytics. On the other hand, to have high levels of analytical data utility, data should

not be changed much from its full disclosure form, which in effect makes privacy breach very likely. Data privacy and analytical utility are therefore inversely proportional, requiring delicate balancing act for any private data release aimed at supporting data analytics to achieve its goal [6].

The differential privacy model, a private data release approach that promises to deliver the balance between privacy and analytical utility, is expressed in a theoretical mathematical format, which is not utilitarian. This has made the model face an implementation challenge that needs to be resolved to allow its adoption and wide-spread application [10][7]. In particular, the choice of an appropriate privacy parameter, the epsilon value, that guarantees the privacy without eroding analytical utility has not been utilitarian in a manner that enables software developer be able to implement the model. There is a need for a practical and heuristic approach on how to arrive at the right value. This is what this model review paper aims to achieve.

This paper is a narrative model review that seeks to operationalize the differential privacy by making the choice of privacy parameter become practical and heuristic, hence making it utilitarian. The method used was to synthesize the available literature that was obtained from journals and other academic materials sourced through online searching. A gap that needs to be filled was then identified and a solution to it is provided.

2. Data analytics & data release

Data analytics allows the examination of data sets with a view of extracting useful information by identifying and analyzing behavior and patterns using both qualitative and quantitative techniques [11]. This, however, poses a threat of disclosure of private information about individuals whose data is in the data sets.

If the data sets are used by the curator for analysis, privacy concerns would not exist. However, the curators may wish to a release the data to third party analyst or to the public, who may perform secondary data analytics on the data and the analyst may need to link to other data sets from other sources for the process of analytics to be successful. Data analytics may call for the need to interact with various data sets, in order to attain the hidden patterns and relationships among the data sets. It is important for curators to know which data to release to a third party or even to the general public to enable further analysis using the dataset.

To provide a conducive environment for data analytics, there is need to enable private data release – releasing anonymized data, whose individuals who are the subject matter, are protected from disclosure to unauthorized parties. Such data should retain analytical utility to allow analysts to draw some insights from them. The released data should not have associations with the individuals, who are the data subject [12]. In this way, the data released protects the privacy of the individuals but retains analytical utility.

Private data release is necessary for both privacy preserving data publishing (PPDP) and privacy preserving data mining/analysis (PPDM/A). In PPDP, the aim is to provide the public with deidentified or synthetic data for further investigation. The purpose of PPDP influences the kind of data to be published. If the intention is just to inform, without further analysis expected, then contingency tables – a matrix format showing statistics of one variable in a row and those of another in a column, and histograms – a plot showing the frequency distribution of data, may be used [13] - [17]. However, publishing data intended to be used for analytical purposes needs to retain reasonable analytical utility. In PPDM/A, the data is not released to the analyst, instead, they are allowed to interact with the data set through aggregated queries [10], [18].

3. Privacy and analytical utility

The need for a mechanism that would enable data release that supports data analytics in an environment that guarantees privacy of the individuals whose data is held without sacrificing its analytical utility is the drive behind numerous research work in the area.

Some of the disclosure control mechanisms that have been used to limit privacy loss, such as the anonymization-based techniques, have been shown to diminish the analytical utility of the data due to the alterations made in attempt to mask the individuals in the data sets [18].

To achieve private data release that supports data analytics, it is necessary to balance the two competing goals: privacy of individuals and analytical utility of the data sets [17], [18]. The two are antagonistic in that very private data will not be of much analytical use (utility), while high data analytical utility implies high accuracy, which is likely to cause a privacy breach [19]. Therefore, the choice of an anonymization mechanism to be used in enhancing private data release must be done with a good trade-off between the two goals [12]. We highlight the two main categorizations of the privacy mechanisms used, namely anonymization and differentially privacy models.

4. Anonymization privacy models

A typical data set has three main types of attributes that describe the subject (individual). 1) Explicit Identifiers (EID) – attributes whose values uniquely (directly) identify an individual in the data set. Such includes name, national identification card number, etc. 2) Quasi Identifiers (QID) – attributes whose values on their own may not be able to identify an individual, but when combined with values of other QID, have potential to identify and individual. Examples include gender, age, etc. 3) Sensitive Attributes (SA) – attributes whose values are confidential in nature and individuals in the data set would be uncomfortable if revealed or associated with them. Such includes income, ailment diagnosis, etc. [20].

Any anonymization (de-identification) technique used must remove, hide, or suppress all the explicit identifiers to make sure individuals are not revealed. However, the QID and SA should remain.[18], [21]

K-Anonymity is one of the techniques used for de-identification where a group of records of the dataset with same attribute values is referred to an equivalence class. This technique requires that each equivalence class in the data set, has at least k (a constant number) members, meaning each member of the equivalence class has k-1 other elements that cannot be distinguished from it. The value of k is a constant whole number, i.e., number of records in a given equivalence class. The technique is known to protect against identity disclosure – being able to identify a record in the dataset, but not attribute disclosure – where, from the attribute values, one is able to learn about a group of records. K-Anonymity is further demonstrated to be susceptible to homogeneity attack and background knowledge attack [22]–[24].

The ℓ-diversity model, which is an improvement of k-anonymity, requires that the values of the sensitive attributes in each equivalence class have at least ℓ (a constant number) well-represented values. This means that the values for the sensitive attributes of a given equivalence class are such that there are ℓ indistinguishable records, where ℓ is greater or equal to two (i.e., $\ell \geq 2$). The model is reported to be prone to skewness attack and similarity attack [22], [23].

The t-closeness model improved the ℓ-diversity model by requiring that the distance between the distribution of sensitive attributes in a class is not more than a threshold t [22].

The three models and their affiliates are commonly referred to as anonymization models, and are known to lack mechanism

to deal with background information the analyst may have, hence unable to provide guarantee of privacy. Another concern of these models is the loss of analytical utility of the data that is caused by alterations of the original data values in order to achieve the masking aspect [6], [21], [25]. This means the two goals, privacy & utility, sought in private data release may not be achieved using these models. This makes the models unsuitable for private data release.

5 Differential privacy model

This is the privacy model that promises to achieve both the goals of privacy and utility of data sets. Indeed, it has become the de facto model in private data release [25]. The model requires that the probability distribution in the released results area essentially remains the same, irrespective of whether an individual's data is included in the dataset or not. This way, the presence or absence of an individual in the data set does not influence the result of the analysis. This ensures that the analyst does not learn about an individual in the dataset, but is able to learn about the population represented by the members in the data set [26]–[28].

In particular, the ε-differential privacy (ε-DP) provides a provable and quantifiable privacy guarantee, as well as a trade-off between the privacy and data utility. An algorithm, in the context of analytics, a query, is said to satisfy ε-DP, if and only if, the difference in probability (Pr) of any query outcome (S) of two data sets (D1 & D2), which differ only by a maximum of one entry, only varies by a factor of exponential (e) to the power of epsilon (ε) [29], [30]. Formally, a randomized algorithm M, is said to provide ε - differential privacy if for all data sets D1 and D2 differing in not more than one record, and all S ⊆ Range(M), then, equation (1) below applies [26].

$$Pr[M(D1) \in S] \leq e^{\varepsilon} * Pr[M(D2) \in S] \quad (1)$$

The ε-DP mechanism ensures that the computational result of dataset does not change significantly due to the inclusion or exclusion of an individual in the dataset. This is achieved by adding carefully calibrated noise to the true results, making the output insensitive to changes in individual record [15], [31].

One important concept for guaranteeing ε-DP is mechanism sensitivity, denoted as Δf, that measures the maximum change in output of a mechanism as a result of change in individual record. The literature has shown that the sensitivity and the epsilon (ε) value determine the noise to be added for a mechanism to satisfy ε-DP. When dealing with real numbers, Dwork et al. [16] proved that noise from the Laplace mechanism with scale of Δf/ε would satisfy ε-DP, while when using integers [32], Ghosh et al. [33] proved that noise from the geometric mechanism with scale of ε/Δf would do the same [15], [34]. Their views are widely supported in the literature [35]. Mechanism sensitivity (Δf) can easily be computed from the data set but there is not much published work on how the value of epsilon (ε) is obtained.

The effectiveness of ε-DP approach is, therefore, very much dependent on the choice of epsilon (ε) value, which is the privacy parameter, also called the privacy budget. The privacy parameter (or budget) controls the trade-off between the privacy guarantee and the data analytical utility. Small epsilon values lead to higher privacy due to more noise added to mask the data, but it also implies less accurate data, hence low utility. Large epsilon values lead to less noise added, meaning high accuracy; hence high utility, but the individuals are at high risk of re-identification [36]. This follows the fundamental law of information recovery, which indicates that very accurate answers to many questions destroy privacy in a big way [27], [37]. Therefore, getting the right value of epsilon (ε) is an important aspect in operationalizing the differential privacy [30].

Despite the promise of ε-DP achieving the two antagonistic goals (privacy & utility) necessary in private data release, its usage/application is reported to be very low [18]. This is attributed to its theoretical mathematical expression that is not easily implemented [10] and, in particular, to its description of the privacy parameter that is not in a utilitarian format that is readily applicable [38]. Derivation of the privacy parameter (the epsilon value), is not heuristic – i.e., not self-explanatory, meaning not everyone can derive it, for a given dataset.

Differential privacy model has been ascertained to be the one that can give the much-sought balance between privacy and data utility in the private data release. Its practical application has, however, been found to be limited, despite its promising potential. The comprehension and interpretation of its theoretical mathematical formulation that would lead to wide application have been stifled by the challenge of establishing the privacy parameter, the epsilon (ε), which is the guarantee of the privacy being provided by the mechanism.

For ε-DP to get widespread application, a practical and heuristic way of determining the privacy parameter needs to be provided and proven empirically, for software developers to know how to apply it. We take a look at attempts that have been made in trying to arrive at the appropriate value of this privacy parameter.

6. Choice of privacy parameter in differential privacy

Choosing the value of epsilon that satisfies ε-DP has been reported not to be a trivial matter; however, there is dearth of research on how to determine it. In some research works [30], the value is picked without explanation of how it was arrived at, or is simply assumed to be a certain value. If an empirical and heuristic method of determining the value is provided, implementation of ε-DP is likely to be embraced and widely used [30], [36].

Two methods for determining the appropriate value of the epsilon that were found in the literature differ in their approach significantly. One by Hsu et al. [30], views the epsilon (ε) as a factor of: 1) Budget of conducting the study (B), 2) Target accuracy or error margin (T), 3) Expected cost of individual participating or not participating in the study (E) and 4) probability error or confidence measure (α). They proposed a formula to obtain the appropriate privacy parameter as the equation (2) below.

$$\varepsilon \leq \ln\left[1 + \frac{BT^2}{12E\ln\frac{3}{\alpha}}\right] \qquad (2)$$

The authors reported that the expected benefits of the participants they studied were in monetary form, which made them quantifiable. Lee and Clifton [36] work used a mathematical formulation approach in their coming up with the formula of computing the value of privacy parameter, the epsilon value. In so doing, they were able to assume certain values of probability and used them to prove their formula. Their approach was theoretical in nature. However, a practical approach method that is validated empirically and is heuristic is needed. A heuristic method would enable analysts and system developers to apply the method to get the probability of re-identification, which is then applied to compute the appropriate privacy parameter for a given dataset on their own.

This probability of re-identification depends on factors that are intrinsic to the dataset as well as external factors that vary from one region to another. The intrinsic factors are 1) the uniqueness – characterizes the amount of unique elements in the dataset, and 2) distinguishing power of each attribute [4], [39]. The external factors are 1) the technical skills and resources available to the analyst and 2) the availability of linkage data sets that can be linked to the anonymous data sets [18]. There is need

therefore to model how to arrive at the re-identification considering these factors. They made two fundamental assumptions that made it possible to operate: 1) that participants were afraid of some bad events and 2) that they were able to estimate their expected cost of these bad events [30].

The challenge with this model is that the parameters must be established at the point of data collection. That is, an analyst who gets data sets that were collected without those parameters stated may not have a way of determining the epsilon value, hence not able to implement the model. Therefore, we observe that this approach may be applicable in some circumstances but not in all situations.

The second method is by Lee and Clifton [36], which views the epsilon (ε) as a factor of: 1) Global sensitivity (Δf), 2) Maximum distance between possible solutions (Δv), 3) Size of data set (n) and 4) Probability of being identified (p). They proposed a formula of getting the appropriate privacy parameter as the equation (3) below.

$$\varepsilon \leq \frac{\Delta f}{\Delta v} \ln \frac{(n-1)p}{1-p} \quad (3)$$

The sensitivity, maximum distance, and dataset size are inherent in the dataset and can be computed or read from the dataset for any given dataset. Once the probability of identification is established, the privacy parameter ε will be known. We find this approach applicable by the analysts, on data sets they collect and those collected by others. Lee and Clifton [36] concluded that a mechanism for establishing the probability of identification is important in making determination of appropriate value of the privacy parameter ε.

7. Modelling reidentification probability

Getting an appropriate estimate of the re-identification probability is critical in computing the appropriate value of epsilon, which is the privacy parameter (the budget). The rest of variables for computing epsilon (ε), i.e. sensitivity, maximum distance, and dataset size, are computed directly from the dataset. The privacy parameter regulates the trade-off between the privacy and utility, as well as the amount of noise to be added so that the output results satisfy ε-DP.

Following factors identified as influencing re-identification [4], [18], [39], the re-identification probability can only be estimated for a given region and for a specific time. This is because factors such as the analytical skills of the analysts and the resources available to them, that were identified as influencing re-identification, will vary from region to another. The same applies to the availability of linkage data sets to the analysts to whom they can refer.

The reviewed literature identified each of the factors influencing re-identification as factor on its own. It is our considered view that the factors do not work in isolation, but instead work together in contributing to re-identification. Therefore, it is necessary to examine their combined cause and effect, and that is what informed the formulation of the proposed model in Fig. 1.

The causal relationship model that is postulated to influence the re-identification probability was arrived at in consideration of the factors stated above. The identified factors are latent variables that need measurement indicators. The appropriate indicators for each factor were identified and, therefore, represent the model as shown in Fig. 1.

8. The proposed model

Fig. 1. represents the model that can be used to estimate the re-identification probability of a given region. The model adopted the Structural Equation Modelling (SEM) for its ability to work with latent variables, also known as a construct. A construct is a representation of factor that cannot be measured directly; instead, its indicators are used to measure it.

There are four constructs that form the structural or inner model.

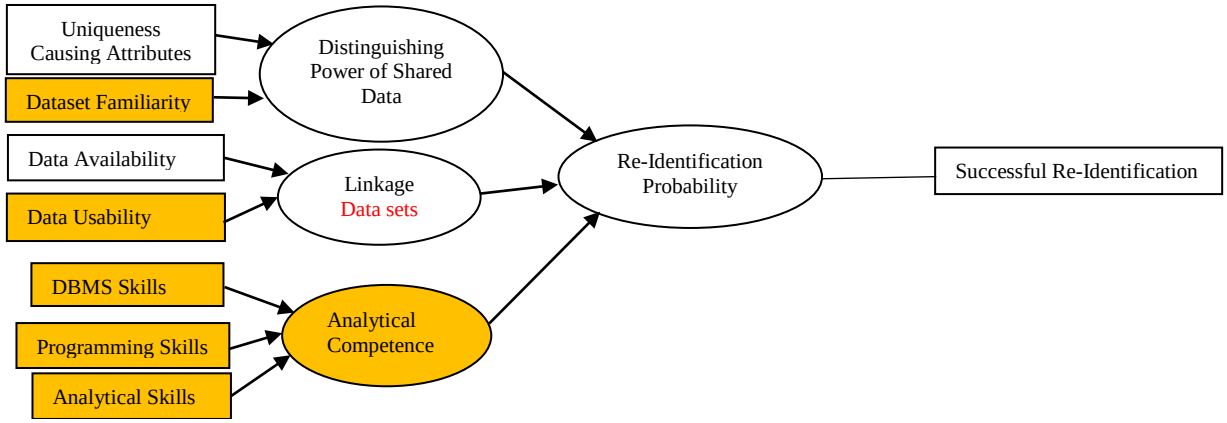

Fig. 1. Proposed Re-Identification Probability Estimation Model

The distinguishing power, linkage data sets, and analytical competence are the independent constructs, also known as exogenous constructs, that predict the dependent construct, re-identification probability, also known as the endogenous construct.

Distinguishing power construct refers to the ability to single out an entity from a dataset, which leads to re-identification of the entity. This is postulated to be determined by the characteristics of the quasi-attributes that are in the data sets and the background information (dataset familiarity) that the analyst may be in possession of. The two become the indicators or measured variables representing the construct. The data set familiarity was not emphasised in the previous model.

The linkage data sets construct refers to the various data sets that the analyst may need to compare with the anonymized data set released by the curator. Such data sets would be containing both explicit and quasi-identifiers. The analyst then matches the quasi-identifiers from the released data to linkage data sets and then uses the explicit identifier to disclose who the entity that had been de-identified is, causing the privacy bleach. This is postulated to be measured using the linkage dataset availability, accessibility, and its usability. Previously, the emphasis was only on the availability, but usability is equally very key.

Analytical competence construct refers to the ability of the analyst interacting with the released data sets to work with data sets and be able to extract relevant data/information aiding in re-identification. The construct was postulated to be measured through analyst's skills in databases, programming statistical mathematics, data mining, and data analytics. This is a new inclusion in the model to emphasise the role of data user in the re-identification process.

The re-identification probability construct refers to the likelihood of an analyst re-identifying an entity that was previously de-identified at the time of data release. The construct is measured by successful re-identification that does happen.

The measured variables (indicators) and the constructs they represent form the measurement or outer model. Our proposed structural equation model hence has the outer and the inner models. Both of them need to be validated empirically for the proposed model to be said to be validated.

9. Conclusions

We have demonstrated the need to come up with a way of determining the probability of being re-identified as the aspect that will make the choice of an appropriate privacy parameter become practical and heuristic. We further demonstrated that the probability of being re-identified will vary from one region to another. This implies that the epsilon value can only therefore be estimated for a given region.

We have improved the model by introducing new indicators for both distinguishing power and linkage data set. We further introduced a

new construct (Analytical Competence) to cover the data user or the analyst.

The proposed model needs to be validated empirically, by collecting data and experimenting it in a given region to get the re-identification probability. Once a region has established this probability of re-identification, the value would be plugged into the formula of determining the privacy parameter, epsilon (ε), as stated in equation (3), as the value of P. In this way, we would succeed in making the choice of privacy parameter practical and heuristic, making the application of the ε-differential privacy utilitarian and hence more applicable.

10. Acknowledgment

We thank Dr. Bonface Ngari Ireri, Multimedia University of Kenya, for moral support and also assisting with the editing work during the writing of the manuscript.

References

[1] R. Gupta, "Journey from Data Mining to Web Mining to Big Data," Int. J. Comput. Trends Technol., vol. 10, no. 1, pp. 18-20, 2014.

[2] M. Schroeck, R. Shockley, J. Smart, D. Romero-Morales, and P. Tufano, "Analytics: The real-world use of big data," IBM Glob. Bus. Serv. Saïd Bus. Sch. Univ. Oxford, pp. 1-20, 2012.

[3] X. Yao, X. Zhou, and J. Ma, "Differential Privacy of Big Data: An Overview," 2016 IEEE 2nd Int. Conf. Big Data Secur. Cloud (BigDataSecurity), IEEE Int. Conf. High Perform. Smart Comput. (HPSC), IEEE Int. Conf. Intell. Data Secur., vol. 9, no. 2, pp. 7-12, 2016.

[4] A. Ramachandran, L. Singh, E. Porter, and F. Nagle, "Exploring re-identification risks in public domains," in 2012 10th Annual International Conference on Privacy, Security and Trust, PST 2012, 2012, pp. 35-42.

[5] A. Cavoukian and D. Reed, "Big Privacy: Bridging Big Data and the Personal Data Ecosystem Through Privacy by Design," 2013.

[6] S. Reddy and O. Prakash, "UTILITY-PRIVACY TRADEOFF IN DATABASES : AN INFORMATION THEORETIC APPROACH," Int. J. Eng. Sci. Res., vol. 4, no. 10, pp. 608-612, 2014.

[7] K. Nissim et al., "Differential Privacy: A Primer for a Non-technical Audience * (Preliminary version)," no. 1237235, 2017.

[8] M. Alfalayleh and L. Brankovic, "Quantifying privacy: A novel entropy-based measure of disclosure risk," in Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 2015, vol. 8986, pp. 24-36.

[9] C. Dwork, "An ad omnia approach to defining and achieving private data analysis," Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), vol. 4890 LNCS, pp. 1-13, 2008.

[10] T. Zhu, G. Li, S. Member, W. Zhou, and P. S. Yu, "Differentially Private Data Publishing and Analysis: A Survey," IEEE Trans. Knowl. Data Eng., vol. 29, no. 8, pp. 1619-1638, 2017.

[11] S. Sruthika and N. Tajunisha, "A Study on Evolution of Data Analytics To Big Data Analytics and Its Research Scope," 2015 Int. Conf. Innov. Information, Embed. Commun. Syst., 2015.

[12] G. Cormode, "The Confounding Problem of Private Data Release," Proc. 18th Int. Conf. Database Theory, pp. 1-12, 2015.

[13] N. Mohammed, R. Chen, B. C. M. Fung, and P. S. Yu, "Differentially private data release for data mining," in Proceedings of the 17th ACM

SIGKDD international conference on Knowledge discovery and data mining - KDD '11, 2011, p. 493.

[14] K. Nissim and A. Smith, "Smooth Sensitivity and Sampling in Private Data Analysis ∗," Proc. thirty-ninth Annu. ACM Symp. Theory Comput., pp. 75-84, 2007.

[15] C. Dwork, F. McSherry, K. Nissim, and A. Smith, "Calibrating Noise to Sensitivity in Private Data Analysis," Proc. 3rd Theory Cryptogr. Conf., pp. 265-284, 2006.

[16] C. Dwork, F. McSherry, K. Nissim, and A. Smith, "Calibrating Noise to Sensitivity in Private Data Analysis," Theory Cryptogr. SE - 14, vol. 3876, pp. 265-284, 2006.

[17] N. Mohammed, R. Chen, B. C. M. Fung, and P. S. Yu, "Differentially private data release for data mining," Proc. 17th ACM SIGKDD Int. Conf. Knowl. Discov. data Min. - KDD '11, p. 493, 2011.

[18] S. L. Garfinkel, "NISTIR 8053 De - Identification of Personal Information NISTIR 8053 De - Identification of Personal Information," 2015.

[19] L. Yin et al., "Re-identification risk versus data utility for aggregated mobility research using mobile phone location data," PLoS One, vol. 10, no. 10, pp. 1-23, 2015.

[20] H. Vaghashia and A. Ganatra, "A Survey: Privacy Preservation Techniques in Data Mining," Int. J. Comput. Appl., vol. 119, no. 4, pp. 20-26, 2015.

[21] A.-E.-E. Abdou Hussien, N. Hamza, and H. A. Hefny, "Attacks on Anonymization-Based Privacy-Preserving: A Survey for Data Mining and Data Publishing," J. Inf. Secur., vol. 4, no. April, pp. 101-112, 2013.

[22] N. Li, T. Li, and S. Venkatasubramania, "t -Closeness : Privacy Beyond k -Anonymity and -Diversity," in IEEE 23rd International Conference, 2007, no. 3, pp. 106-115.

[23] A. Machanavajjhala, D. Kifer, J. Gehrhe, and M. VENKITASUBRAMANIAM, "L-Diversity : Privacy Beyond k - Anonymity," in Proceedings of the 22nd International Conference on Data Engineering, 2006, vol. V, pp. 1-36.

[24] L. Sweeny, "k- ANONYMITY: A MODEL FOR PROTECTING PRIVACY 1," Int. J. Uncertainty, Fuzziness Knowledge-Based Syst., vol. 10, no. 5, pp. 557-570, 2002.

[25] H. H. Nguyen, J. Kim, and Y. Kim, "Differential Privacy in Practice," J. Comput. Sci. Eng., vol. 7, no. 3, pp. 177-186, 2013.

[26] C. Dwork, "Differential privacy in new settings," Proc. Twenty-First Annu. ACM-SIAM Symp. Discret. Algorithms, pp. 174-183, 2010.

[27] C. Dwork and A. Roth, "The Algorithmic Foundations of Differential Privacy," Found. Trends® Theor. Comput. Sci., vol. 9, no. 3-4, pp. 211-407, 2013.

[28] C. Dwork, "A firm foundation for private data analysis," Commun. ACM, vol. 54, no. 1, p. 86, 2011.

[29] H. Ebadi, D. Sands, and G. Schneider, "Differential Privacy: Now it's Getting Personal," Proc. 42nd Annu. ACM SIGPLAN-SIGACT Symp. Princ. Program. Lang. - POPL '15, pp. 69-81, 2015.

[30] J. Hsu et al., "Differential Privacy: An Estimation Theory-Based Method for Choosing Epsilon," Proc. 2014 IEEE 27th Comput. Secur. Found. Symp., pp. 398-410, 2014.

[31] X. Cheng, S. Su, S. Xu, P. Tang, and Z. Li, "Differentially private frequent sequence mining," IEEE Trans. Knowl. Data Eng., vol. 28, no. 11, pp. 2910-2926, 2016.

[32] L. Fan and H. Jin, "A Practical Framework for Privacy-Preserving Data Analytics," Www'15, pp. 311-321, 2015.

[33] A. Ghosh, T. Roughgarden, and M. Sundararajan, "Universally Utility-Maximizing Privacy Mechanisms," SIAM J. Comput., vol. 41, no. 6, pp. 1673-1693, 2012.

[34] A. Ghosh, T. Roughgarden, and M. Sunararajan, "Universally utility-maximizing privacy mechanisms," Proc. 41th STOC, vol. 41, no. 6, pp. 1673-1693, 2009.

[35] R. Sarathy and K. Muralidhar, "Evaluating laplace noise addition to satisfy differential privacy for numeric data," Trans. Data Priv., vol. 4, no. 1, pp. 1-17, 2011.

[36] J. Lee and C. Clifton, "How much is enough? Choosing Epsilon for differential privacy," in Information Security, 14th International Conference, ISC 2011, 2011, vol. 7001 LNCS, pp. 325-340.

[37] C. MIT, "Big Data Privacy Workshop: Advancing the State of the Art in Technology and Practice," Cambridge, Massachusetts, 2014.

[38] P. Mohan, A. Thakurta, E. Shi, D. Song, and D. E. Culler, "GUPT: Privacy Preserving Data Analysis Made Easy," Sigmod, p. 12, 2012.

[39] F. K. Dankar, K. El Emam, A. Neisa, and T. Roffey, "Estimating the re-identification risk of clinical data sets.," BMC Med Inf. Decis Mak, vol. 12, no. September 2009, p. 66, 2012.

Peter N. MUTURI graduated from School of Computing and Informatics, University of Nairobi, Kenya, with MSc. in Information Systems in 2010. He is now a Computer Science PhD candidate, at the School of Computing and Informatics, University of Nairobi. He is currently a Lecturer, Computer Science Department, Faculty of Computing and Information Technology at Multimedia University of Kenya. His domains of work are: Big Data, Data Analytics, and Data Privacy.

Andrew M. KAHONGE graduated from University of Birmingham with MSc. Advanced Computer Science, Specializing in Distributed Systems and Networks, Neural Computation and Virtual Reality, in 2003. He Attained his Doctoral degree from School of Computing and Informatics, University of Nairobi, in Computer Science in 2013, Specializing in Web Security and User Behavior Modeling. At present he is a Senior Lecturer in the School of Computing and Informatics, University of Nairobi. He has published several papers in reputable and international conferences and journals.

Christopher K. CHEPKEN is a Senior Lecturer at the School of Computing an Informatics, University of Nairobi where he has served since 2004. He holds a PhD in Computer Science from the University of Cape Town, South Africa (2013) and a Master of Applied Computer Science from the University of Nairobi (2006), where he also obtained his Bachelor's degree in Computer Science (2004). Christopher has published widely in the area of Applied Computing and other related topics. He has also mentored and supervised several masters and PhD students to completion. His specialization areas include ICT for Development, Systems security and Software Engineering.