Identity Lock – Privacy-Preserving Data Publishing Tool

Neha Maroof Siddiqui[1] Sara Benish[2] Tehreem Qamar[3]

Abstract

In today's world, data sharing is very common. Currently, the strong movement is occurring towards publishing data for statistical studies. In this case, data publishers are providing some sort of data to the research field, but they do not know what kind of things the 3rd party can do with the data provided to them. Data preservation is an important aspect when sharing data because attackers can easily disclose a person's identity and their personal information. Hence, in order to secure privacy, different methodologies are implemented on data. This paper presents Identity Lock - the Privacy Preserving Data Publishing (PPDP) tool, uses various anonymization techniques and implements k-Incognito, l-Incognito and ε-Differential Privacy algorithm to hide and anonymize data. The software also performs the experimental evaluation in order to calculate the performance of the algorithm on the basis of how much utility and privacy is maintained.

Keyword: Privacy Preserving, PPDP, Data Preservation, Anonymization, Privacy Models

1 Introduction

Demand of microdata is becoming diverse. Organizations collect and share this microdata for knowledge-based decision making [1]. In statistics, microdata is a set of records containing personal information [2]. It contains some personal information that causes privacy issues. Not only that, study shows that 87% of United States population was identified easily from published datasets [3].

Statistical studies like enumeration, population factors, health statistics and road accidents records, all created from data. While publishing data, privacy concern and preservation is considered as an imperative factor for viable use of data. This information is kept in electronic configuration, without causing any trouble to a person [3]. The consistently increasing velocity of data makes security a challenging task, particularly when the data is high in storage.

In 2006, Netflix an online DVD-rental company released their data to improve its movie recommendation algorithm [4]. The company released anonymized data, but just 16 days later two specialists from The University of Texas easily distinguished clients by coordinating the informational data from other sites like IMDb. On December 17, 2009, four Netflix clients documented a legal claim against Netflix, asserting that Netflix had disregarded U.S. reasonable exchange laws and the Video Privacy Protection Act by releasing the datasets [5]. Another case of sensitive information leakage occurred when AOL (American Online)- an online service provider released their search log of 657,000 American citizens from which an individual

[1] *Jinnah University for Women, Karachi | nehamaroofsiddiqui@gmail.com*
[2] *Jinnah University for Women, Karachi | sarabenish@yahoo.com*
[3] *Jinnah University for Women, Karachi | tehreem_qamar10@hotmail.com*

named Thelma Arnold was identified. Later searched data was removed from the websites but the damage is already done [6].

Generally, privacy concerns are identified with validation, data accessing, data encryption, and data publishing. Numerous data holders distribute the microdata of their organization for various purposes such that released data don't violate a person's privacy [4]. To address these serious privacy violations, data should be published after certain anonymization processes are applied to it. The research area focusing this issue is known as PPDP (Privacy-Preserving Data Publishing). It is an important step for securely publishing microdata for research analysis and statistical studies. Until now, different methods [5],[6],[7] are proposed that mainly focused on protecting the disclosure of private information, while providing the utility in published data.

The contribution of this paper can be summarized as:

- Survey of different anonymization algorithm and their limitations in preserving privacy and providing utility data.
- Development of a tool named "Identity Lock" that implement the different algorithm (i.e. k-Incognito, l- Incognito and ε-Differential Privacy)
- Evaluation of algorithms on a variety of dataset in order to keep the privacy of each individual and provide utility data for further statistical need.

The rest of the paper is organized as follows: Section II presented the background of PPDP extended by some general privacy techniques and survey of related work on different privacy models. The proposed system is presented in Section III with the experimental analysis in Section IV, while the conclusion discussed the final verdict of the research study in Section V.

2 Background & Related Work

A Privacy-Preserving Data Publishing

The approach of analyzing and acting upon data is extremely important for various organizations [8]. The sharing of data may lead to misuse or excessive data distortion. Privacy-Preserving Data Publishing is a concept providing method for publishing useful information while preserving individual's privacy. Figure 1 described the process of publishing privacy preserved data by following steps presented as:

- An owner collects raw data from their organizations.
- Anonymization techniques are applied to preserve data privacy.
- Once the privacy is preserved, data is released to publish for research and statistical analysis.

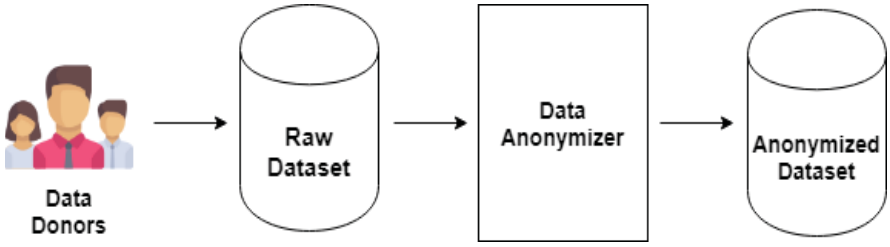

Figure 1: Privacy Preserving Data Publishing Architecture.

For the process of anonymization, microdata is generally categorized into three forms:

- Direct identifier: There are some attributes that easily identify a person's identity such as name, address, user ID, etc.
- Quasi Identifier (QI): The group of attributes which helps in recognizing a person such as class, age or gender.
- Sensitive Attribute: The data fields that contain an individual's personal information such as disease, salary.

B General Techniques

The most general techniques use to anonymize data are:

1) Generalization: It is a process that transforms the group of records into more generalized one. It removes direct identifiers from the datasets and then assigns a common value to the group of data records that possess the same kind of data [9]. It is one of the flexible technique [10] but it lacks in providing data utility [11].

2) Bucketization: It aims to preserve privacy by dividing records according to a quasi-identifier and assigns a unique ID to each division[12]. Then, both the quasi-identifier and sensitive value in records are published separately. By applying this process, specific values are not lost, but it breaks the relation of QI and sensitive attribute [13].

3) Suppression: Similar to generalization, suppression first removes the direct identifiers, then changes specific values of quasi-identifier (QI) to *, completely hides some values indicating that replaced record is not meant to be published [14]. The replacement of values with "*" causes information loss which is the main drawback of this method [15].

4) Perturbation: This technique is based on randomization [16]. It can be implemented by replacing the original value with any random value. The perturbed data records change sensitive values while quasi-identifier remains unchanged due to which resultant dataset does not ensure privacy protection [15].

5) Slicing: It is done on records by dividing datasets horizontally and vertically[17]. Vertical partitioning grouped co-relative attribute in a column and horizontal

partitioning grouped sets of records in buckets. Each bucket is then randomly permutated.

Table 1: General Techniques

Techniques	Advantages	Disadvantages
Generalization	It protects identity disclosure by replacing specific information.	It is prone to homogeneity attack and the background knowledge attack.
Bucketization	It prevents dataset from record linkage attack by assuming a clear separation in between QIs and SAs.	It failed to prevent membership disclosure i.e. doesn't protect attribute disclosure to sufficient extent.
Suppression	It provides identity disclosure risk by suppressing the real value.	"*" value disturbed utility at high rate, especially during statistical analysis.
Perturbation	The attacker cannot perform the sensitive linkages or recover sensitive information from the published data.	The perturbation approach does not provide a clear understanding of the level of indistinguishability of different records.
Slicing	It preserves better utility than that of generalization while protecting privacy.	Due to the correlation of high attribute, privacy violation may happen in slicing technique.

C Related Work

K-anonymity was first proposed by Sweeney [18] in which she replaces values in the dataset by less-specific value. After which, the Datafly approach [19] uses a heuristic to perform generalization on quasi-attributes. However, again, no formal foundations or abstractions have been provided. Samarti's work [20] uses k-minimal generalization but failed to maintain optimal minimum information loss. The study in [20], [21] on the method is known as minimal generalization which is independent of the purpose of released data. Another approximation algorithm was proposed by El-Amawy[22] which provides optimal anonymization. However, the method failed when larger values of k are desired. Moreover, [6] presented a method which uses a clustering mechanism but results in minimizing the utility of data as suppression hide most of the information.

Fung et al. in his work [23] discussed the main goal of the data release by implementing classification, resulting in k-anonymize data that is optimal and minimizes the cost metric. Generally, achieving the optimal k-anonymity is NP-hard [22], [24].Besides the general anonymization techniques, LeFevre et al. Studies an extension of k-anonymization [25] and proposed the multidimensional k-anonymity. Moreover, LeFevre et al. [26] broadened the previously stated multidimensional approach for anonymizing a specific task i.e. classification. Xu et al. [27] discussed different greedy approaches to use k-anonymization for cell generalization. His work proved that the anonymized microdata results in less utility loss than that of used by LeFevre et al. [26].

Apart from k- anonymization, l-diversity [28] save both the data privacy and its utility. It obtains anonymization with the diversity of sensitive values on quasi-identifying groups. In Bucketization [29], publishing sensitive values separately from the quasi-identifier, secure linking attacks and maintain data utility as no changes are made on specific values. But still, the identity of the victim can easily be known by the attacker as it does not provide membership disclosure [30]. Further Li et al. [31], discusses the limitations of l-diversity and introduced a t-closeness technique to overcome privacy attacks. The t-closeness [32] calculates Earth Mover Distance (EMD) between two distributions for all the field contained in the dataset and a sensitive attribute. In t-closeness, there is a correlation between sensitive attributes of a dataset and QIDs; t-closeness degrades utility of privacy preserved data. Secondly, for sensitive numerical data, the t-closeness is unable to prevent attribute linkage attacks [33]. Thirdly, t-closeness uses EMD measurement that is not perfect and flexible enough to impose different privacy levels on different the sensitive attributes.

Perturbation used in [34], added noise to datasets. Xiao [35] and Chaytor [36] state that perturbation is only suitable in cases that focus on preserving privacy while false records don't affect research analysis result. Chaytor and Wang [36], work on randomly assigning sensitive attribute by dividing its domain. It results in high error where small ranges are used. Following the randomization and perturbation, Laplacian noise is included in differential privacy [37] to improve the sensitivity of data.

The general loss metric is presented in [38] uses a generalized data to calculates a normalized loss of information. Dewri et al. [39], use a weighted k-anonymity, focusing on the privacy-utility issue for better results.

Table 2: Summary of Privacy Models

Authors	Description
Sweeney [12]	K-anonymity, presented the protection model to anonymize data implementing generalization and replacing values in dataset by less-specific value. But the approach used lacks in securing individual's privacy.
Sweeney [17]	Data fly, applied heuristic approach to perform generalization on quasi-attributes. It guarantees k-anonymous transformation but doesn't provide the minimal generalization.
Samarti, Sweeney [18], [19]	Minimal generalization, implemented generalization and suppression on datasets but generalize data more than its needed and failed in providing optimal minimum information loss.
P. Kulasinghe [20]	The approximation algorithm, proposed to provide optimal anonymization. However, method failed when larger values of k are desired.
G. Aggarwal [21]	Clustering mechanism, maintain k-anonymity. However, this approach failed to maintain the utility of data, as suppression hide most of the information useful for data mining or research work.

Fung et al. [22]	Classification, presented a top-down approach to iteratively refine the data from a general state into a special state. However, the leakage of two data points can result in complete disclosure of information.
LeFevre et al. [24]	Mondrian Multi-dimensional anonymity, where generalization is performed on multi-dimensional data using approximation algorithm. But it lacks in providing as much data utility as needed for research studies
Xu et al. [26]	Greedy approximation algorithm, to use k-anonymization for cell generalization. His work proved that the anonymized micro data results in less utility loss.
Ercan Nergiz [27]	δ-presence, published sensitive values separately from quasi identifier, secure linking attacks and maintain data utility.
Li et al. [29],	T-closeness, discusses the disadvantages and limitations of l-diversity and then introduced t-closeness technique to overcome various privacy attacks. However, this technique does preserve feature disclosure but identity is still disclosed.
Xiao et al [33]	Optimal perturbation, achieves anonymization with a multi-level perturbation approach that release multiple data sets anonymized on different levels of privacy.
Chaytor and Wang [34]	Small domain generalization, work on randomly assigning sensitive attribute by dividing its domain. This approach retains more data yet not guaranteed information disclosure risk.

3 Identity Lock

Identity Lock is designed to implement k-Incognito [40], l-Incognito [41] and ε-Differential Privacy [42]. We have chosen these algorithms on the basis of the following facts:

- Chosen algorithms are extensively cited.
- These algorithms use different strategies that work on both categorical and numerical attributes.

It also evaluates the algorithm on the basis of utility and privacy maintained by the algorithms. Figure 2 describes the software design of Identity Lock. The software requests a file of supported format from a user. As soon as user uploads the file to be anonymized, the software reads data from the uploaded file, display it in GUI and ask a user to choose sensitive attribute (SA) and a privacy model to proceed further. Once the required parameters are selected, the software starts anonymizing the given dataset by following the privacy model selected by the user.

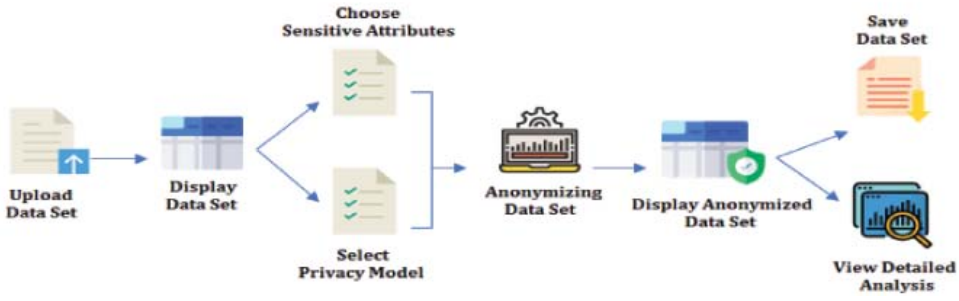

Figure 2: Processing Model of Identity Lock.

Identity Lock supports the following algorithms: k-Incognito, l-Incognito, and ε-Differential Privacy

- K-Incognito uses bottom-up search to anonymize the quasi attributes in datasets. It works on protecting identity disclosure. Yet the privacy is vulnerable when attacker have strong background knowledge of individual [40].
- To overcome such problems, l-Incognito diversify the values of sensitive attributes within an equivalence class. It protects datasets against attribute disclosure [41].
- For statistical data, ε-Differential Privacy uses additive noise approach to ensures privacy [42].
- Levenshtein metric is implemented to measure utility maintained by anonymized dataset [43].
- Shannon Entropy metric is used to measure the privacy of anonymized dataset [44].
- To compare execution time of an algorithm, the software monitored time from start till the end of the process.

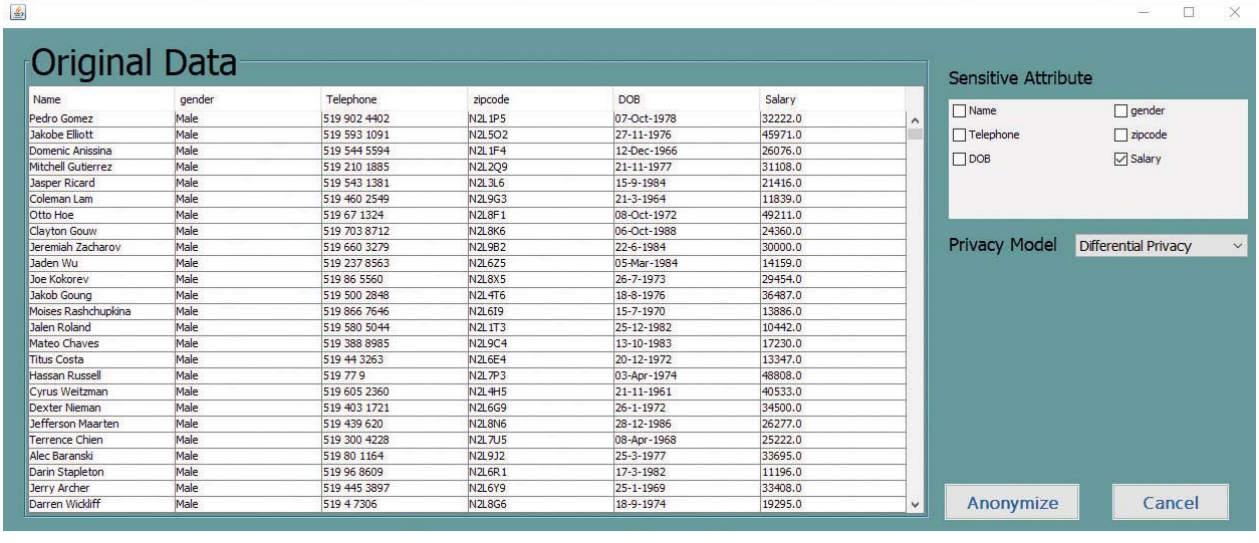

Figure 3: Uploaded Dataset

Identity Lock also featured the automatic anonymization of the dataset based on the privacy model which maintained good utility-privacy of data. This feature implements all the algorithms, compares utility and privacy maintained by each algorithm and then display result that best guarantees both the utility of data and privacy of the individual's in a dataset. The anonymized data can then be saved in .csv or .xls file format. The user can also view a detailed analysis of utility and privacy maintained by the privacy model and the time taken by the system to anonymize the whole dataset. The detailed mechanism of algorithms is discussed below:

A K-Incognito

K-Incognito technique follows the global-recording model called a full-domain generalization. DGH (Domain Generalization Hierarchy) [40] is formed for each QI attributes Q. The number of valid domain generalization[45] varies by the depth of its VGH (Value Generalization Hierarchy). For a dataset containing multiple QI attributes, the domain generalization hierarchies of each individual attribute are merged to build a generalization lattice.

Incognito algorithm[7] uses bottom-up search to pass over this generalization lattice. It starts by checking single-dimensional nodes of QI attributes, proceeds by iterating to an increasing subgroup of QI sets in the lattice to check k-anonymity requirement. If a node fulfils the property of k-anonymity, then all of its direct generalizations is marked, guaranteeing that they also satisfy the property of k-anonymity. The algorithm terminates when all the combinations of QI attributes have been considered.

This method may be in-efficient with respect to time but the anonymized dataset contains the maximum quantity of information that makes this algorithm an optimal solution[45] for preserving privacy.

B ε-Differential Privacy:

Differential privacy is considered as "State-of-Art" technique in the data privacy field. Noise addition technique was first outlined by Dwork [42] to address the anonymization of statistical data. It guarantees privacy by utilizing noise addition perturbation methods that transform sensitive attribute by adding calculative noise to it. The differential privacy method ensures that an attacker can't succeed in misusing information about any person in the dataset.

According to Dwork [46], if two datasets D1 and D2 differ or disagree in a single record, an anonymized algorithm A is said to satisfy ε-differential privacy, if it results R supports the equation:

$$\frac{P[A(D_1) \epsilon R]}{P[A(D_2) \epsilon R]} \leq e^{\epsilon} \qquad (1)$$

Where P represents the probability of an event occur and refers to the statistical distance to determine the strength of privacy. It has been noted by C.Dwork[42] that smaller values for ε give more privacy while ε= 0 is said to be completely differential private. However, utility risk increases with smaller ε value.

C L-Incognito

K-anonymity privacy definition is vulnerable to adversaries that have strong background knowledge of individuals represented in the dataset. [41] l-diversity tries to overcome such vulnerabilities by diversifying sensitive values within each equivalence class.

The toolbox implements an approach to multiple-sensitive attributes relies on the solution described by Sweeney [9]. All sensitive values should merge into a single attribute, which then specified as the only sensitive attribute. Incognito anonymization with k-anonymity as the privacy definition allowed suppression by default. This suppression approach does not apply to l-diversity since the purpose of anonymization is to diversify the sensitive value distribution. Therefore, suppression is disabled within this version of Incognito.

D Experiment Subject

1) Time Efficiency: Execution time is one of the aspects that must be considered while processing anonymization. To compare the execution time between algorithms, the software monitored time from the start till the very end of the anonymization process

2) Utility Metric: Due to the insufficiency of standardized metrics, it is challenging to measure data utility. Some metrics as discussed in [47], [48] are suitable for utility measure of numeric data but not provide any mechanism to measure utility for categorical data.For our evaluation criteria, the level of information remained in dataset after the completion of the anonymization process that is measure by using string metric proposed in [32]. In this analysis, 1.0 is considered as the best utility score, whereas zero measured as the worst score for maintaining utility. The Levenshtein metric measures the similarity between two words by calculating an edit distance. The distance is the number of deletions, insertions, or substitutions required to transform the anonymized data to original data. The procedure for calculating the Levenshtein distance between two strings X of length m and Y of length n is to calculate step by step in a matrix of order m x n edit distance between different sub-strings of X and Y. The corresponding values are stored in the matrix up to the box (m,n) which expresses the minimum distance between X and Y.

3) Privacy Metric: Attacker aims to re-identify anonymized data by linking with an individual's data record. The protection model applied in software used to measure identity disclosure risk by applying Shannon's entropy [34]. In this analysis, 1.0 is considered as best utility score, whereas zero is measured as the worst score for maintaining utility. Shannon's entropy metric was proposed in [35] as a measure of effective anonymity set. In order to measure identity disclosure risk of an anonymized dataset, Shannon entropy metric calculates the probability for the occurrence of values in each attribute. Usually, the more uncertain the probability, the less the disclosure risk. Shannon's entropy can be used to estimate this uncertainty, by applying:

$$H(X) = -\sum_{i=1}^{n} P(x)_i \log_b P(x)_i \qquad (2)$$

where is a calculated probability of an attribute, compute it as the proportion of attribute in the dataset. The above equation quantifies the degree of utility contained in the dataset. The higher the entropy value, the less the disclosure risk.

4 Experiment's Results

For our experimental analysis we used following dataset:

- Employee's Salary dataset that consists of around 2000 records and 6 attributes (Name, Telephone, Age, Sex & Salary).
- Crime dataset that consists of around 1100 records and 7 attributes (Name, Block, Gender, Race, Age, Case Number and Cause of Incident).
- Marriage dataset that consists of around 800 records and 5 attributes (Country, Sex, Education, Marital Status & Age at Marriage).
- Disease dataset that consists of around 2000 records and 6 attributes (Name, Telephone, Age, Sex, Marital Status & Disease).
- Energy Consumption Dataset that consists of around 2000 records and 4 attributes (Residence Address, Zip code, Occupation & Energy Consumed).

In this section, we present the result generated from the datasets by presenting the comparison methodology.

A Employee Salary Dataset

Figure 4 shows that the algorithms perform efficiently with respect to the execution time. However, the execution time for l-Incognito is comparatively high.

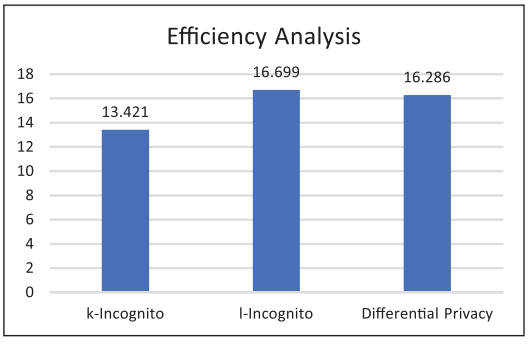

Figure 4: Efficiency Analysis of Employee Salary Dataset.

From Figure 5, it is obvious that K-Incognito performs better in maintaining utility. On the other hand, ε-differential privacy scored worst as it uses an additive noise mechanism for the anonymization process.

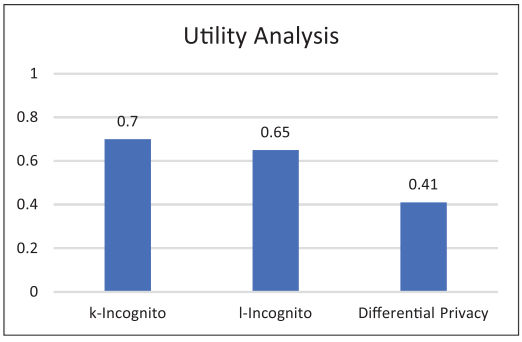

Figure 5: Utility Analysis of Employee Salary Dataset.

Figure 6 presents the result related to privacy analysis of the privacy models. K-Incognito leads to minimize the risk disclosure and as the information is anonymized enough, it provides higher data privacy results.

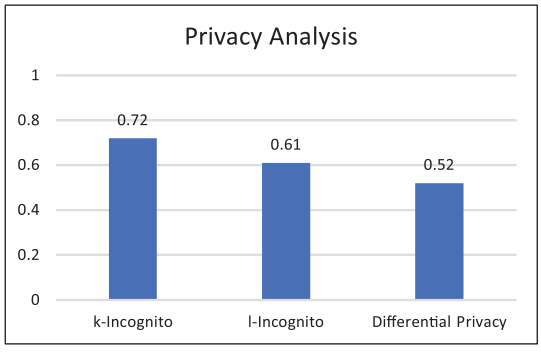

Figure 6: Privacy Analysis of Employee Salary Dataset.

B Patient Disease Dataset

Figure 7 shows that the execution time for l-Incognito scored the worst where as, k-Incognito complete the whole process is little less time as compared to the other algorithms.

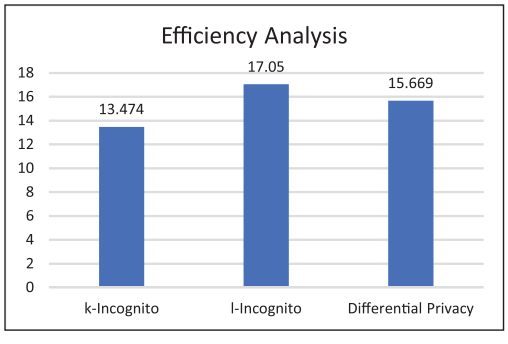

Figure 7: Efficiency Analysis of Patient Disease Dataset.

As in Figure 8, it is observed that again K-Incognito performed better in maintaining utility and differential privacy scored worst to gain utility of anonymized data.

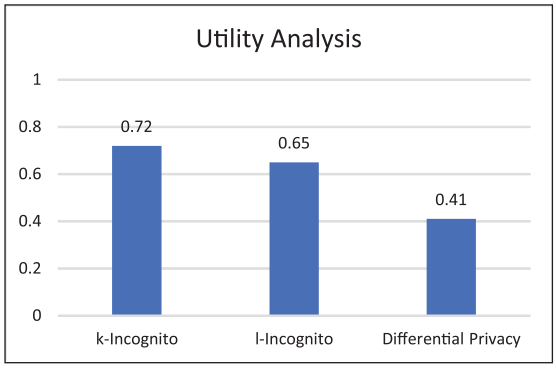

Figure 8: Utility Analysis of Patient Disease Dataset.

Figure 9 presents the privacy score of the anonymized dataset. K-Incognito leads to minimize the risk disclosure and as the information is anonymized enough, it provides higher data privacy results.

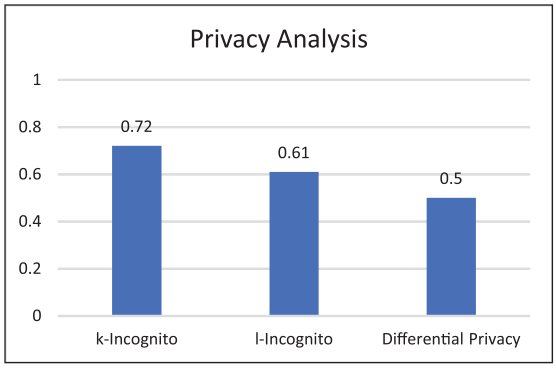

Figure 9: Privacy Analysis of Patient Disease Dataset.

C Crime Incident Dataset

Figure 10 shows that the algorithms perform the process in much less time as the number of quasi-attributes is less. The execution time for l-Incognito scores best while ε-differential privacy takes more time to complete anonymization.

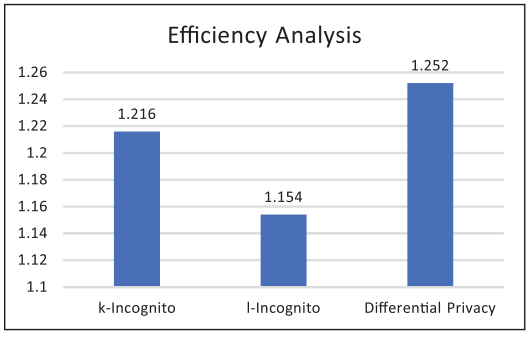

Figure 10: Efficiency Analysis of Crime Incident Dataset.

Figure 11, shows that K-Incognito scored best in maintaining utility. On the other hand, ε-differential privacy scored worst because of its additive noise mechanism for the anonymization.

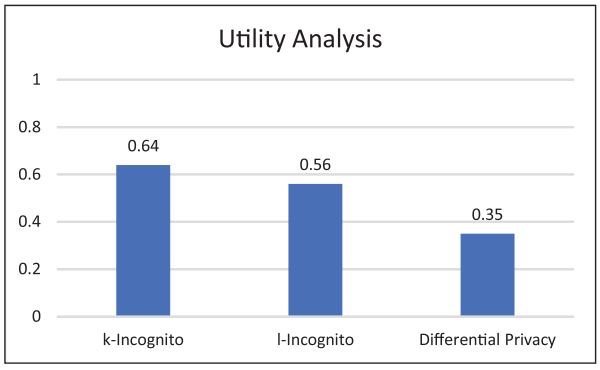

Figure 11: Utility Analysis of Crime Incident Dataset.

Figure 12 gives the result of privacy analysis of Crime Incident Dataset. K-Incognito reduces the risk disclosure and as the information is anonymized enough; it provides higher data privacy results.

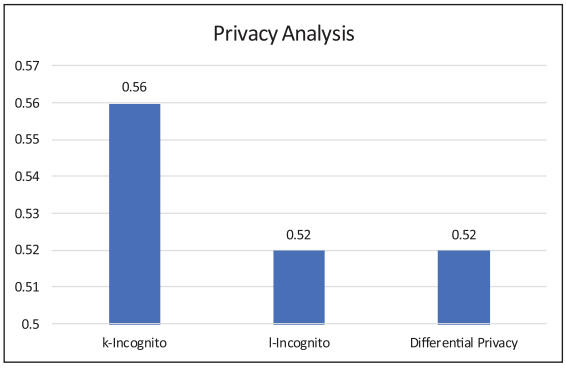

Figure 12: Privacy Analysis of Crime Incident Dataset.

D Marriage Dataset

Figure 13 shows that ε-differential privacy takes much more time while other algorithms perform the whole process in much less time. K-Incognito scored the best with respect to execution time.

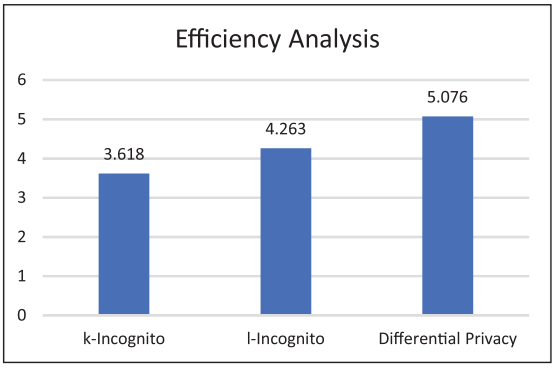

Figure 13: Efficiency Analysis of Marriage Dataset.

From Figure 14, can be found that K-Incognito performed most efficiently in maintaining the utility of dataset whereas, ε-differential privacy scored worst.

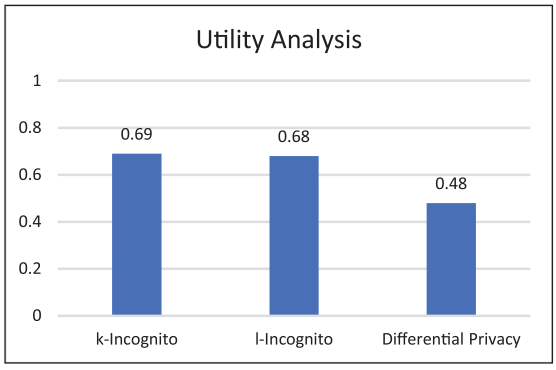

Figure 14: Utility Analysis of Marriage Dataset.

Figure 15 shows that the algorithms failed to provide the best privacy. l-Incognito provides better results related to privacy. However, k-Incognito is the worst performer when it comes to Marriage dataset.

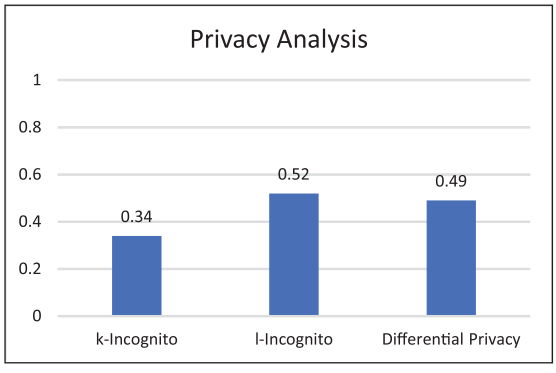

Figure 15: Privacy Analysis of Marriage Dataset.

E Energy Consumption Dataset

As this data consists of only 2 quasi identifiers, the algorithm takes a few seconds to anonymize the data. Figure 16 shows that the execution time for l-Incognito is comparatively less as compared to that of other algorithms.

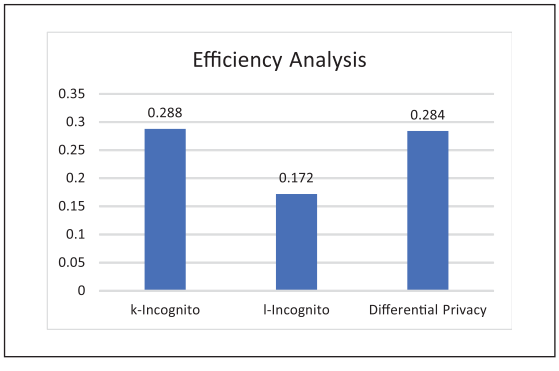

Figure 16: Efficiency Analysis of Energy Consumption Dataset.

From Figure 17, it is obvious that K-Incognito maintains better utility. Whereas, ε-differential privacy and l-Incognito provide less utility for anonymized data.

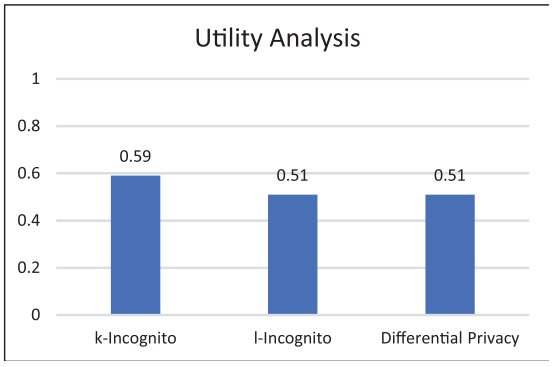

Figure 17: Utility Analysis of Energy Consumption Dataset.

Figure 18 presents the result related to privacy analysis of the privacy models. K-Incognito leads to minimize the risk disclosure and as the information is anonymized enough, it provides higher data privacy results.

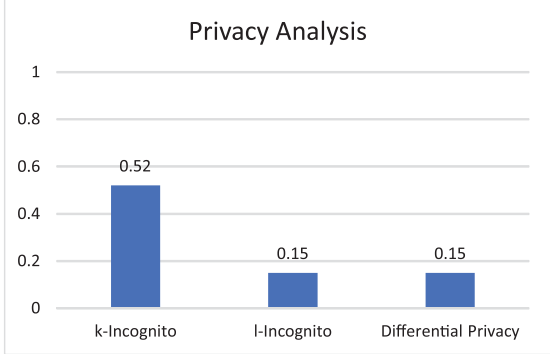

Figure 18: Privacy Analysis of Energy Consumption Dataset.

5 Conclusion

In this paper, we developed a data-publishing tool that preserves both the privacy and utility of data before sharing data to the external world. The software implemented different privacy model, including k-Incognito, l-Incognito and differential privacy. It evaluated the performance of algorithm using datasets related to different domains and compared the results in terms of execution time, data utility and data privacy. The software also features an automatic anonymization mode that provides results based on the algorithm that maintained utility-privacy of data at a higher rate. As per future enhancement, the tool can implement other algorithms. We can also introduce other supported file formats and work on the parameters i.e. de-identification risk of anonymized data.

Using k-Incognito, l-Incognito and differential privacy, we come to this conclusion that the execution time of algorithm depends upon the no. of quasi-identifiers contained by the dataset. The algorithm performs more efficiently when dataset consists of a smaller number of

quasi-identifier. Moving forward, k-Incognito performs better in maintaining both utility and privacy in most of the cases. However, l-Incognito provides better privacy results when the dataset contains discrete values.

References

[1] U. Trivellato, "Microdata for Social Sciences and Policy Evaluation as a Public Good," no. 11092, 2017.

[2] K. Wang, Privacy Preserving Data Publishing, vol. 42file:///, no. 4. 2010.

[3] L. Candela, D. Castelli, P. Manghi, and S. Callaghan, "On research data publishing," Int. J. Digit. Libr., vol. 18, no. 2, pp. 73–75, 2017.

[4] A. Narayanan and V. Shmatikov, "How To Break Anonymity of the Netflix Prize Dataset," 2006.

[5] O. Gkountouna, "A Survey on Privacy Preservation Methods," pp. 1–30, 2011.

[6] G. Aggarwal et al., "Achieving anonymity via clustering," ACM Trans. Algorithms, vol. 6, no. 3, pp. 1–19, 2010.

[7] V. Ciriani, S. De Capitani di Vimercati, S. Foresti, and P. Samarati, "k-Anonymous data mining: a survey," Adv. Database Syst., pp. 105–136, 2008.

[8] H. R. Asmaa and B. M. Y. Norizan, "Privacy preserving data publishing: Review," Int. J. Phys. Sci., vol. 10, no. 7, pp. 239–247, 2015.

[9] L. Sweeney, "Simple demographics often identify people uniquely," Carnegie Mellon Univ. Data Priv. Work. Pap. 3. Pittsburgh 2000, pp. 1–34, 2000.

[10] D. Dubli and D. K. Yadav, "Secure Techniques of Data Anonymization for Privacy Preservation," vol. 8, no. 5, pp. 2015–2018, 2017.

[11] G. S and V. P, "A Survey on Privacy Preserving Data Publishing," Int. J. Cybern. Informatics, vol. 3, no. 1, pp. 1–8, 2014.

[12] D. O. F. Informatics, "Survey of Privacy-Preserving Data Publishing Methods and Speedy : a multi-threaded algorithm preserving ? -anonymity Βιβλιογραφική επισκόπηση μεθόδων προστασίας της ιδιωτικότητας δεδομένων προς δημοσίευση και Speedy : ένας πολυνηματικός αλγόριθμος που δι," no. October, 2015.

[13] B. C. M. Fung, K. Wang, R. Chen, and P. S. Yu, "Privacy-preserving data publishing," ACM Comput. Surv., vol. 42, no. 4, pp. 1–53, 2010.

[14] Y. Xu, T. Ma, M. Tang, and W. Tian, "A survey of privacy preserving data publishing using generalization and suppression," Appl. Math. Inf. Sci., vol. 8, no. 3, pp. 1103–1116, 2014.

[15] "GUIDE TO BASIC DATA ANONYMISATION TECHNIQUES Published 25 January 2018," no. January, 2018.

[16] A. Antoniades et al., "The effects of applying cell-suppression and perturbation to aggregated genetic data," IEEE 12th Int. Conf. Bioinforma. Bioeng. BIBE 2012, no. November, pp. 644–649, 2012.

[17] T. Li, N. Li, J. Zhang, and I. Molloy, "Slicing: A New Approach to Privacy Preserving Data Publishing," IEEE Trans. Knowl. Data Eng., vol. PP, no. 99, p. 1, 2009.

[18] L. SWEENEY, "k-ANONYMITY: A MODEL FOR PROTECTING PRIVACY," Int. J. Uncertainty, Fuzziness Knowledge-Based Syst., vol. 10, no. 05, pp. 557–570, 2002.

[19] L. Sweeney, "Datafly: a system for providing anonymity in medical data," pp. 356–381, 1998.

[20] Pierangela Samarati, "Protecting respondents' identities in micro- data release," IEEE Trans. Knowl. Data Eng., vol. 13, no. 6, pp. 1010–1027, 2001.

[21] L. Sweeney, "ACHIEVING fc-ANONYMITY PRIVACY PROTECTION USING GENERALIZATION AND SUPPRESSION," Int. J. Uncertain., vol. 10, no. 5, pp. 571–588, 2002.

[22] P. Kulasinghe and A. El-Amawy, "On the Complexity of Optimal Bused Interconnections," IEEE Trans. Comput., vol. 44, no. 10, pp. 1248–1251, 1995.

[23] B. C. M. Fung, K. Wang, and P. S. Yu, "Anonymizing Classification Data for Privacy Preservation," IEEE Trans. Knowl. Data Eng., vol. 19, no. 5, pp. 1–14, 2007.

[24] G. Aggarwal et al., "Anonymizing tables," Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), vol. 3363 LNCS, pp. 246–258, 2005.

[25] K. LeFevre, D. J. DeWitt, and R. Ramakrishnan, "Mondrian multidimensional K-anonymity," Proc. - Int. Conf. Data Eng., vol. 2006, p. 25, 2006.

[26] K. Lefevre and D. J. Dewitt, "Workload-Aware Anonymization," pp. 277–286, 2006.

[27] J. Xu, W. Wang, J. Pei, X. Wang, B. Shi, and A. W.-C. Fu, "Utility-based anonymization using local recoding," Proc. 12th ACM SIGKDD Int. Conf. Knowl. Discov. data Min. - KDD '06, p. 785, 2006.

[28] J. Gehrke, "ℓ -Diversity : Privacy Beyond k -Anonymity," vol. V, pp. 1–47, 2000.

[29] M. Ercan Nergiz, M. Atzori, and C. W. Clifton, Hiding the Presence of Individuals from Shared Databases. 2007.

[30] A. Paul Singh and D. Parihar Asst, "A Review of Privacy Preserving Data Publishing Technique," 2013.

[31] N. Li, T. Li, and S. Venkatasubramanian, "t-Closeness: Privacy Beyond k-Anonymity and-Diversity."

[32] Y. Sei, H. Okumura, T. Takenouchi, and A. Ohsuga, "Anonymization of Sensitive Quasi-Identifiers for l-diversity and t-closeness," IEEE Trans. Dependable Secur. Comput., pp. 1–1, 2017.

[33] J. Li, Y. Tao, and X. Xiao, Preservation of Proximity Privacy in Publishing Numerical Sensitive Data. 2008.

[34] R. Agrawal and R. Srikant, "Privacy-preserving data mining," ACM SIGMOD Rec., vol. 29, no. 2, pp. 439–450, Jun. 2000.

[35] X. Xiao, Y. Tao, M. Chen, and A. Kumar Maji, "Optimal Random Perturbation at Multiple Privacy Levels."

[36] R. Chaytor and K. Wang, "Small Domain Randomization: Same Privacy, More Utility," 2150.

[37] C. Dwork, F. Mcsherry, K. Nissim, and A. Smith, "Calibrating Noise to Sensitivity in Private Data Analysis."

[38] V. S. Iyengar, "Transforming data to satisfy privacy constraints," in Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '02, 2002, p. 279.

[39] R. Dewri, I. Ray, I. Ray, and D. Whitley, "k-Anonymization in the Presence of Publisher Preferences," IEEE Trans. Knowl. Data Eng., vol. 23, no. 11, pp. 1678–1690, Nov. 2011.

[40] K. LeFevre, D. J. D. J. DeWitt, and R. Ramakrishnan, "Incognito: efficient full-domain K-anonymity," SIGMOD '05 Proc. 2005 ACM SIGMOD Int. Conf. Manag. data, pp. 49–60, 2005.

[41] A. Machanavajjhala, J. Gehrke, D. Kifer, and M. Venkitasubramaniam, "ℓ-Diversity: Privacy beyond k-anonymity," Proc. - Int. Conf. Data Eng., vol. 2006, p. 24, 2006.

[42] C. Dwork, "Differential Privacy."

[43] M. Tabata, Y. Hosokawa, O. Watanabe, and J. Sohma, "Direct Evidence for Main Chain Scissions of Polymers in Solution Caused by High Speed Stirring," Polym J, vol. 18, no. 10, pp. 699–712, 1986.

[44] "Shannon entropy."

[45] M. S. Simi, K. S. Nayaki, and M. S. Elayidom, "An Extensive Study on Data Anonymization Algorithms Based on K-Anonymity," IOP Conf. Ser. Mater. Sci. Eng., vol. 225, no. 1, 2017.

[46] C. Dwork, A. Roth, C. Dwork, and A. Roth, "The Algorithmic Foundations of Differential Privacy," Found. Trends R □ Theor. Comput. Sci., vol. 9, pp. 211–407, 2014.

[47] D. Sinwar, R. Kaushik, and M. Tech Scholar, "Study of Euclidean and Manhattan Distance Metrics using Simple K-Means Clustering," www.ijraset.com, vol. 2, 2014.

[48] F. Rahutomo, T. Kitasuka, and M. Aritsugi, "Semantic Cosine Similarity."