PeerJ Computer Science

Improved angelization technique against background knowledge attack for 1:M microdata

Rabeeha Fazal[1], Razaullah Khan[2], Adeel Anjum[3], Madiha Haider Syed[3], Abid Khan[4] and Semeen Rehman[5]

[1] Department of Computer Science, COMSATS Institute of Information Technology, Islamabad, Pakistan
[2] Department of Computer Science, University of Engineering and Technology, Mardan, Pakistan
[3] Institute of Information Technology, Quaid-i-Azam University, Islamabad, Pakistan
[4] College of Science and Engineer, University of Derby, Derby, United Kingdom
[5] Institute of Computer Technology, Technische Universität Wien, Wien, Austria

ABSTRACT

With the advent of modern information systems, sharing Electronic Health Records (EHRs) with different organizations for better medical treatment, and analysis is beneficial for both academic as well as for business development. However, an individual's personal privacy is a big concern because of the trust issue across organizations. At the same time, the utility of the shared data that is required for its favorable use is also important. Studies show that plenty of conventional work is available where an individual has only one record in a dataset (1:1 dataset), which is not the case in many applications. In a more realistic form, an individual may have more than one record in a dataset (1:M). In this article, we highlight the high utility loss and inapplicability for the 1:M dataset of the θ-Sensitive k-Anonymity privacy model. The high utility loss and low data privacy of (p,l)-angelization, and (k,l)-diversity for the 1:M dataset. As a mitigation solution, we propose an improved (θ^*,k)-utility algorithm to preserve enhanced privacy and utility of the anonymized 1:M dataset. Experiments on the real-world dataset reveal that the proposed approach outperforms its counterpart, in terms of utility and privacy for the 1:M dataset.

Subjects Security and Privacy, Internet of Things
Keywords Security, Privacy, Internet of Things (IoT), Anonymity

INTRODUCTION

Modern technologies, such as the Internet of Things (IoT) and Big Data are the key enablers to revolutionizing today's modern society in different fields, for example, the Electronic Health Records (EHRs) (*Dang et al., 2019*; *Muftuoglu, Kızrak & Yildirim, 2022*; *Amin et al., 2022*). The government or private organizations collect the EHRs *via* the IoT devices and share them for further statistical analysis and policymaking (*Moonsamy & Singh, 2022*; *Sheikhtaheri et al., 2022*). However, the EHRs belong to an individual, these are very confidential and crucial in the medical information control system. Sharing such data without privacy implementation is unlawful, because of the possibility of privacy breach and misuse of data (*Dang et al., 2019*; *Sun et al., 2018*; *Al-Khafajiy et al., 2019*; *Al-Khafajiy et al., 2018*; *Fazal et al., 2022*), as shown in Fig. 1, where an attacker compromises the

Submitted 19 April 2022
Accepted 24 January 2023
Published 15 March 2023

Corresponding authors
Madiha Haider Syed,
madiha@qau.edu.pk
Semeen Rehman,
semeen.rehman@tuwien.ac.at

Academic editor
Leandros Maglaras

Additional Information and Declarations can be found on page 32

DOI **10.7717/peerj-cs.1255**

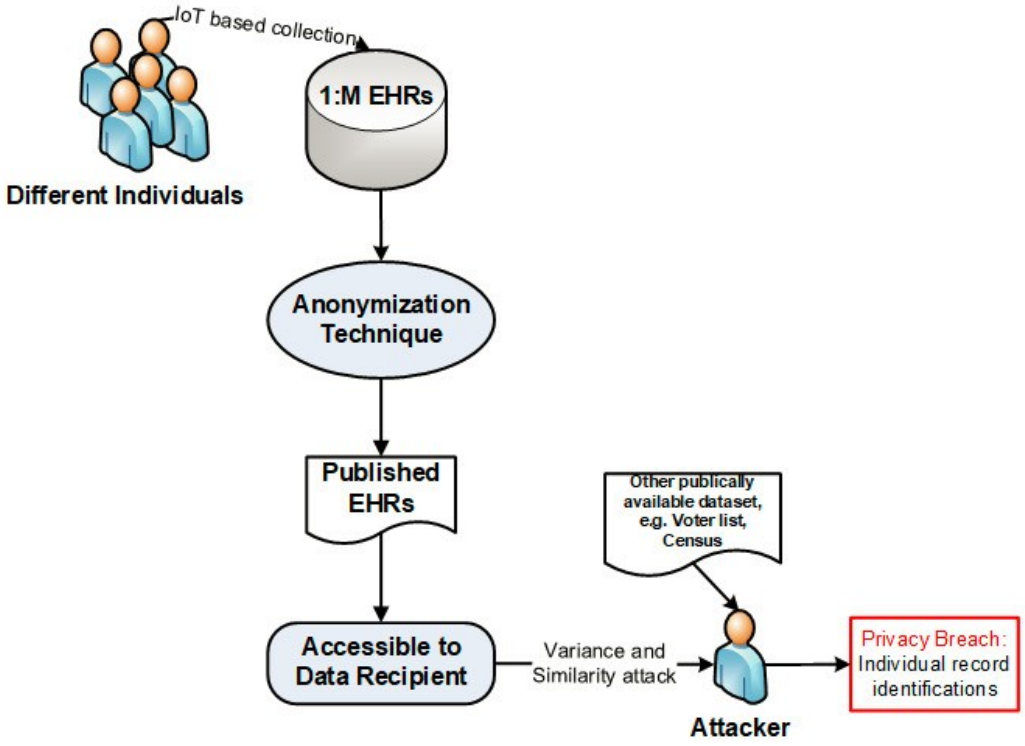

Figure 1 **An example of attacker model.**

Full-size DOI: 10.7717/peerjcs.1255/fig-1

privacy of an individual. Almost all previous anonymization techniques deal with the classical type of data where one person has only one record, *i.e.*, 1:1 dataset; this may not be applicable for all tasks, *e.g.*, health complicated analysis (*Jayapradha et al., 2022*), etc. In a real-world scenario, an individual may have multiple health data records in a dataset or have multiple datasets, known as 1:M microdata (*Gong et al., 2017*).

This article studies the state-of-the-art θ-Sensitive k-Anonymity (*Khan et al., 2020a*), (p, l)-angelization (*Kanwal et al., 2019*), and (k, l)-diversity (*Gong et al., 2017*) algorithms, to highlight respectively their inapplicability for 1:M microdata, privacy breach, and the low data utility. The θ-Sensitive k-Anonymity (*Khan et al., 2020a*) is a numerical measure of privacy strength. The θ denoted the threshold for achieving diversity for sensitive attributes. The θ threshold value is used to achieve diversity through the variance in a group of records (*i.e.*, Equivalence Class - EC). It anonymizes the 1:1 type of data, however, is not applicable for privacy implementation in 1:M microdata, because of the variance calculation inside an EC where each disease value belongs to a specific individual record. If there are more than one disease values (*i.e.*, 1:M data) then the proposed variance calculation method of θ-Sensitive k-Anonymity (*Khan et al., 2020a*) fails. Also, θ-Sensitive k-Anonymity (*Khan et al., 2020a*) algorithm has not proposed any suitable technique for improving the utility of the data.

In the (k, l)-diversity, k is used to protect quasi identifiers (*e.g.*, age, gender, zipcode), and l for the sensitive attributes (*e.g.*, disease, salary) in an EC, so (k, l)-diversity is used

to protect sensitive and quasi attributes. But (k, l)-diversity (*Gong et al., 2017*) partial generalizes the SA values, which causes the sensitive vertical attack (sVer) for the attacker to breach the privacy of data (see Scenario I and Definition 5 for detail). Another privacy focused approach; is the (p, l)-angelization, here the (p, l) shows the extension of the model of p-sensitive k-anonymity (*Ye, Yang and Liu, Yu and Wang, Chi and Lv, Dapeng and Feng, Jianhua, 2009*) and the angelization is used to splits tables into the quasi and sensitive table. The(p, l)-angelization (*Kanwal et al., 2019*) for 1:M-MSA data (*i.e.,* an individual having multiple records and multiple sensitive or confidential attributes), has very low utility considerations. Although, the(p, l)-angelization (*Kanwal et al., 2019*) creates two tables linking through a bucket id (BID). However, that linkage is useless and has no contribution in utility improvement because of the one-to-one correspondence between the two tables. The initial work for 1:M in (k, l)-diversity partially generalizes the SA, giving a privacy leakage window to the adversary and revealing the current medical status of a person with the help of some background knowledge.

In this article, we propose (θ^*, k)-utility technique for 1:M microdata. The (θ^*, k)-utility algorithm overcomes the limitations of (k, l)-diversity, θ-Sensitive k-Anonymity, and (p, l)-angelization by improving data utility and its applicability for 1:M microdata (*Gong et al., 2017*; *Khan et al., 2020a*; *Kanwal et al., 2019*). The motivation subsection discusses in detail the limitations in (k, l)-diversity, θ-Sensitive k-Anonymity and in (p, l)-angelization approaches (*Gong et al., 2017*; *Khan et al., 2020a*; *Kanwal et al., 2019*).

Motivation

Most of the published work in data privacy focuses on either privacy or utility. Keeping a balance between them has always remained an open research problem among researchers. Therefore, the main focus of this research is to come up with an optimized solution where the privacy of the sensitive attributes does not affect the utility of the quasi attribute. A novel solution is needed to independently publish the sensitive table and quasi data without affecting the utility and privacy of the anonymized data. Also, the data should be real-life realistic data, *i.e.,* 1:M data.

The Table 1 of EHRs having 1:M microdata of different individuals are classified into; unique identifier attributes—ID (*e.g.*, name, national identification number, passport number), quasi identifier attributes—QI (*e.g.*, age, gender, zip code, country, religion), and confidential or sensitive attributes—SA (*e.g.*, Symptoms).

Before publishing the data, removing only the ID is not enough because the attacker (*i.e.,* intruder) that can re-identify individual record respondents using some background knowledge—BK (*i.e.,* external source of knowledge that helps in re-identification of an individual) by combining the certain pattern of QIs with some externally available data (*e.g.*, census or voting data), to perform the *linking attack*, (see Fig. 1).

The θ-Sensitive k-Anonymity is a state-of-the-art privacy algorithm, which is applicable to preserve privacy for a single sensitive (SA). However, in addition to low data utility, it cannot work directly for any other type of data. Similarly, the (p, l)-angelization (*Kanwal et al., 2019*), does not focus for data utility improvement. The following scenario explains

Table 1 Original microdata table T.

Patient Record ID	Tuple ID	Name	Age	Zipcode	Gender	Symptoms
p1	t1	Susan	40	2139	Female	Flu
	t2	Susan	40	2139	Female	Chills
	t3	Susan	40	2139	Female	Fever
p2	t4	Ronald	45	2545	Male	Difficult Breathing
	t5	Ronald	45	2545	Male	Lungs Infection
p3	t6	Keran	45	2238	Female	Cough
	t7	Keran	45	2238	Female	Chest Pain
p4	t8	Heather	38	2843	Male	Stomach Pain
	t9	Heather	38	2843	Male	Lungs Infection
p5	t10	Cytnthia	42	2341	Female	Headache
	t11	Cytnthia	42	2341	Female	Tiredness
	t12	Cytnthia	42	2341	Female	Flu
p6	t13	Peter	40	2548	Male	Headache
	t14	Peter	40	2548	Male	Flu
p7	t15	Helan	45	2544	Female	No symptoms
p8	t16	Jonas	32	2538	Male	Headache

θ-Sensitive k-Anonymity approach inapplicability for 1:M microdata and its low data utility, and also (p, l)-angelization (*Kanwal et al., 2019*) high utility loss.

- *Scenario I* Sensitive Vertical Attack (sVer):

The initial work for 1:M in (k, l)-diversity (*Gong et al., 2017*) partially generalize the SA, which give a privacy leakage window to the adversary and reveal the current medical status of a person with the help of some background knowledge; information about a most recent visit or the first visit (either his disease is progressing (worsening), or recovering).

The (k, l)-diversity used a transaction generalization technique (*Gong et al., 2017*) to apply k-anonymity on SAs. The lowest common cut on transaction generalization hierarchy, leaves the single SA from different sub-set unprotected and vulnerable, as shown in Fig. 2. In Fig. 2, if a patient's SAs of multiple records are headache, flu, tiredness, and chills, the (k, l)-diversity will make a lowest common cut on transaction generalization hierarchy and generalize the SAs into 'Low, Chills'. The Low node covers headache, flu, tiredness leaf-nodes and Chills cover chills; which remains 'unprotected and vulnerable and also keeps the Cough leaf-node isolated, which any potential adversary can exploit both the Chills and Cough leaf-nodes, this causes sVer attack, due to partially generalized sensitive attributes.

- *Scenario II* inapplicability for 1:M microdata:

The θ-Sensitive k-Anonymity algorithm (*Khan et al., 2020a*) is based on the θ-threshold value; which is a multiplied component of *variance*, and *observation1*. The variance calculation in each EC (*i.e.*, group of anonymous records) is concerning with the frequency distribution of SA values.

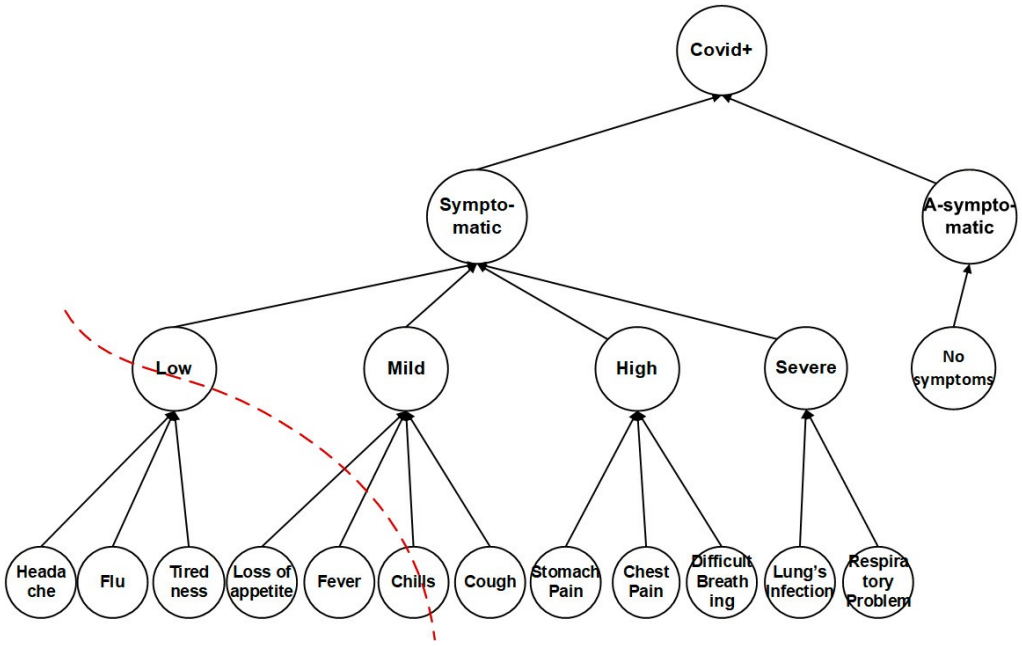

Figure 2 **Sensitive attributes hierarchical structure.**
Full-size DOI: 10.7717/peerjcs.1255/fig-2

If there are more than one SA values for a single record (1:M microdata), then the variance calculation in an EC is not possible. For example to check the inapplicability of θ-Sensitive k-Anonymity for 1:M microdata, consider Table 1 having 1:M microdata. For an anonymized EC, the θ is obtained as Eq. (2), based on variance is obtained as Eq. (1). That shows inapplicability of the threshold value for the 1:M dataset.

$$\sigma^2 = \left(\frac{\Sigma FX^2}{N} - \left(\frac{\Sigma FX}{N}\right)^2\right) \tag{1}$$

$$\theta = Variance\ of\ a\ fully\ diverse\ EC\ (\sigma^2) \times Observation1(\mu). \tag{2}$$

It is obvious that the sensitive values and the corresponding frequency distributions can not be applied to 1:M microdata in Table 1. The reason is that the Symptoms attribute in Table 1 has more than one sensitive value for a specific individual record, which does not represent a frequency distribution for a specific sensitive value. Similarly, if Table 1 is transformed to Table 2; having a complete record in a single tuple. Again, the SA values are more than one in a single tuple, and the θ-Sensitive k-Anonymity approach can not be directly applied to implement diversity in an EC. So, the variance-based θ threshold calculation to obtain a diverse EC is not applicable for 1:M microdata.

- *Scenario III* High utility loss:

  It is based on two approaches.

  (a) The θ-Sensitive k-Anonymity (*Khan et al., 2020a*) approach only focuses the attribute disclosure prevention and does not have any consideration for improving the utility of the

Table 2 θ-Sensitive k-Anonymity inapplicability for 1:M microdata.

Patient record ID	Name	Age	Zipcode	Gender	Symptoms
p1	Susan	40	2139	Female	<Flu, Chills, Fever >
p2	Ronald	45	2545	Male	<Difficult Breathing, Lungs Infection >
p3	Keran	45	2238	Female	<Chest pain, Cough >
p4	Heather	38	2843	Male	<Stomach pain, Lungs Infection >
p5	Cytnthia	42	2341	Female	<Flu, Tiredness, Headache >
p6	Peter	40	2548	Male	<Flu, Headache >
p7	Helan	45	2544	Female	<No Symptoms >
p8	Jonas	32	2538	Male	<Headache >

anonymized data. The algorithm for the θ-Sensitive k-Anonymity begins with checking the k for its minimum value and no further utility improvement consideration is available in the rest of the algorithm.

(b) The (p, l)-angelization (*Kanwal et al., 2019*) splits the microdata table T into Quasi Table (QT) and Sensitive Table (ST). However, the sensitive buckets inside each table have a one-to-one correspondence with one another, which affects the utility in the QT.

Contribution

The main contributions of the proposed (θ^*, k)-utility privacy algorithm are as follows.

- The proposed (θ^*, k)-utility privacy algorithm, categorizes the SA values of 1:M microdata into Low, Mild, High, Severe, and A-symptomatic values, based on the category Table 3 to reshape the original microdata Table 1 into Table 4, for the purpose to get the 1:1 microdata. The SA 1:M record values are replaced with category table SA values. If the SA values are repeated in more than one category, the higher category value is considered and ignored the lower one and stored in history table.
- The proposed algorithm using the angelization approach, anonymizes the microdata T in Table 1 into QT and ST (see Section 5) and are linked through the Bucket ID (BID) using the one-to-many correspondence (*i.e.,* QS-Loose Linkability) for improving utility and privacy, instead of one-to-one correspondence.
- Based of the above points, the experiment results demonstrate the out performance of the proposed (θ^*, k)-utility privacy algorithm, as compared to its counterparts in terms of utility and privacy.

The rest of the article is organized as follows. Section 2 discusses Related Work. Section 3 covers the Preliminaries, and Section 4 discusses the HLPN analysis of previous models. Section 5 discusses the proposed algorithm. Section 6 discusses the Experimental Analysis. Section 7 depicts some Discussion on the base and current proposed work. Section 8 concludes the article with possible future research directions.

Table 3 Category table - CtgT.

ID	Category	Symptoms
1	A-symptomatic	No Symptoms
2	Low	Flu, Headache, Tiredness
3	Mild	Loss of Appetite, Cough, Chills, Fever
4	High	Stomach Pain, Difficult Breathing, Chest Pain
5	Severe	Lung's Infection, Respiratory Problem

Table 4 Original microdata with categorical sensitive values.

Patient record ID	Tuple ID	Name	Age	Zipcode	Gender	Symptoms
p1	t1	Susan	40	2139	Female	Low
	t2	Susan	40	2139	Female	Mild
	t3	Susan	40	2139	Female	Mild
p2	t4	Ronald	45	2545	Male	High
	t5	Ronald	45	2545	Male	Severe
p3	t6	Keran	45	2238	Female	Mild
	t7	Keran	45	2238	Female	High
p4	t8	Heather	38	2843	Male	High
	t9	Heather	38	2843	Male	Severe
p5	t10	Cytnthia	42	2341	Female	Low
	t11	Cytnthia	42	2341	Female	Low
	t12	Cytnthia	42	2341	Female	Low
p6	t13	Peter	40	2548	Male	Low
	t14	Peter	40	2548	Male	Low
p7	t15	Helan	45	2544	Female	A-symptomatic
p8	t16	Jonas	32	2538	Male	Low

RELATED WORK

We studied and analyzed different existing methods and approaches in the literature. Individual privacy is guaranteed with anonymized data since encryption cannot be used for publicly available data to preserve data privacy (*Shahzad et al., 2018*; *Michalas, 2019*). The consensus is before publishing data anonymize it to achieve data privacy. The reason is that, with the anonymized data an individual privacy is preserved. We organize the anonymize techniques into 1:1 and 1:M microdata.

Privacy for 1:1 microdata

Lv & Piccialli (2021) approach is based on the combination of k-anonymity and algorithm of K-A-DP to preserve data privacy. It reduces risks of privacy loss, but It only focuses on numeric values. The author discussed that DP is used for protection when stored in a different place and applied to stored data. It also provides data privacy of electronic health records that cause utility loss (*Choudhury et al., 2019*). *Tu et al. (2018)* discussed the model used for preserving the vulnerability of records using the k anonymity. It is

used to minimize vulnerability to identify the records, but identifying sensitive values can not completely protect that cause breach. *Liu & Li (2018)* proposed an approach that is k-anonymous based on clustering. This process is time-consuming in terms of finding anonymous equivalence.

The article (*Nasir et al., 2017*) prevents the attribute disclosure with skewness attack, which extended the distribution scheme based on the weighted table. It provides low data privacy for the same sensitive values due to non-generalized quasi values. *Lee & Lee (2017)* proposed a model that used the identification factors to predict the re-identification of quasi attributes. The identification probability is based on some factors. But also, it can not fully minimize the aspect of re-identification. *Majeed, Ullah & Lee (2017)* proposed the protection of personal identity information from vulnerability. It provides data privacy to minimize vulnerable records but still has low diversity. *Yaseen et al., (2018)* proposed a model based on conventional, divisor, and cardinality hierarchy. That generates generalization hierarchies but does not focus on textual values. *Anjum et al. (2018)* extended the p+-sensitive k-anonymity and improved this in a balanced p+-sensitive k-anonymity model. It has low diversity for sensitive attributes.

Raju, Seetaramanath & Rao (2019) proposed a model based on slicing, which correlated quasi attributes. The suppression of sensitive attributes depends on the threshold used to create one sensitive value used for all sensitive attributes. A lot of QI and SA suppression may cause huge utility losses. *Song et al. (2019)* proposed a model that used k-anonymous data using noise addition and randomization for categorical data. It may be used for privacy but not use for long-range numeric data. The author presents the improved k-Anonymity with l diversity; the k anonymity and l diversity protect the identity disclosure. It provides data utility (*Jain, Gyanchandani & Khare, 2020*). The θ-Sensitive k-Anonymity is a variance-based θ threshold calculation to obtain diverse Equal Classes. Although the θ-Sensitive k-Anonymity prevents attribute disclosure with sensitive variance and similarity, it cannot be used directly for 1:M microdata (*Khan et al., 2020a*).

Privacy for 1:M microdata

The 1:M type of data is a more realistic form of records stored in EHRs. The(k, l)-diversity model based on 1:M generalization, which prevented attribute disclosure (*Gong et al., 2017*). But it provides low utility and privacy for data. *Wang et al. (2019b)* the algorithm is based on clustering, various decision functions used, but is vulnerable to sensitive attributes. It provides utility for quasi attributes but still has vulnerable to sensitive attributes. The author discussed the approach based on providing privacy for vulnerability disclosure using the model of 1: M MSA- (p, l)-diversity. The attribute disclosure is prevented through this model but also has low data utility due to one-to-one linkage (*Kanwal et al., 2021*). *Anjum et al. (2018)* proposed a heuristic approach to protect the sensitive and quasi values using splitting. It provides privacy but gives low data utility due to 1:1 correspondence. The l-anatomy focuses on the utility of data (*Anjum et al., 2019*). But they publish the sensitive attributes without generalization. This approach deals with achieving utility, but that may not provide enough security. The generic 1:M data privacy (G-model) model uses the

signature method to achieve privacy (*Albulayhi, Tošić & Sheldon, 2020*). However, it has a limitation in which attackers can attack using the signature values to spot records.

The author focuses on COVID-patient data, where the privacy keeps through spatial k-anonymity. But it has limitations to achieving privacy, causing loss of privacy (*Iyer et al., 2021*). This author discussed the approach for privacy-preserving using the k-Anonymity. The focus is using the values of (p, and l) as a threshold. But it cannot prevent the sensitive variance attack, although this approach uses k-anonymity to improve the privacy. It provides low data privacy (*Zhang et al., 2017*). The author used the approach of (α, k) to protect privacy. The Poker dataset is used to measure results, and it cannot protect from the attack to sensitive attributes. It protects from identity disclosure. The attacker accesses the data using a background knowledge attack with low data utility (*Wang et al., 2019a*). The (p, l)-angelization (*Kanwal et al., 2019*) for 1:M-MSA data has very low utility considerations. Although, (p, l)-angelization creates two tables linking through a bucket id (BID). However, that linkage is useless and contributes to utility improvement because of the one-to-one correspondence between the two tables.

The previous approaches do not provide privacy for sensitive attributes and data utility. If someone deals with they cannot directly be used for the 1:M dataset or cannot prevent from background knowledge attack on sensitive attributes for 1:M COVID-19 data (*Gong et al., 2017*; *Kanwal et al., 2019*); so to preserve the 1:M dataset privacy and data utility, we overcame these limitations and extended the θ-Sensitive k-Anonymity because the θ-Sensitive k-Anonymity deals with similarity and variance attack for 1:1 and overcome the limitation of privacy and utility in 1:M microdata to preserve the COVID-19 patient data privacy.

PRELIMINARIES

Let the input table T = {ID, QI, SA} (*i.e.,* Table 1), having 1:M microdata. The $t_i \in$ T is a tuple that represents an individual i having complete or partial record details, depends on the number of tuples that belongs to an individual i. The t_i is a component combination of ID =$\{ A_1^{id}, A_2^{id}, A_3^{id}, \cdots, A_n^{id}\}$, QI =$\{ A_1^{qi}, A_2^{qi}, A_3^{qi}, \cdots, A_n^{qi}\}$, and SA = $\{ A_1^{sa}, A_2^{sa}, A_3^{sa}, \cdots, A_n^{sa}\}$. In the anonymized form the A^{id}s are removed, and a suitable anonymization technique is applied on A^{qi} and A^{sa}, to prevent the identity and attribute disclosures. However, such anonymization techniques should be strong enough to prevent any possible attack; *e.g.*, membership or non-membership attack (ma or nma), sensitive variance attack (sva), categorical similarity attack (csa), sensitive vertical attack (sVer) or any other background knowledge attack (bka). The Table 5 summarizes all the notations used in this article.

The speculation is an adversary, which can access information of any individual using some background knowledge. The adversary model is represented as below.

QT = $\{ A_1^{qi}, A_2^{qi}, A_3^{qi}, \cdots, A_n^{qi}\}$, QT contains all QI attributes and Bucket-ID (BID).

ST = $\{ A_1^{sa}, A_2^{sa}, A_3^{sa}, \cdots, A_n^{sa}\}$, ST consists of the SAs and Bucket-ID (BID), that is linked to the BID in QT.

BID: An identifier between the QT and ST, which links the buckets in both tables through the one-to-many correspondence, known as QS-loose linkability.

Table 5 Summary of notations.

Symbols	Description
T	1:M Microdata Table
k	k-anonymity
EC	Equivalence Class
Γ	T populated with CtgT
ECs	Equivalence Classes
t_i^{sa}	Generalized Sensitive Attributes
A^{id}	Explicit Attributes
BK	Background Knowledge
A^{qi}	Quasi Attributes
A^{sa}	Sensitive Attributes
D_n^{sa}	Set of distinct sensitive attributes
d_i^{sa}	Distinct SA values in dataset
τ	Individual tuple from transform Table $\mathscr{T}$
$\hbar$	Individual history tuple from history table $\mathbb{H}$
sbt	Sensitive buckets
QT	QT contains all QI attributes with QID
ST	ST consists of (Bucket-ID, SAs)
sVer	Sensitive Vertical Attack
CtgT^{cat}	CtgT generalized SA
t_i^{sa}	Sensitive attribute tuples from 1:M
r_i	Single record having may rows
genData	QT
t_i	Tuples from 1:M belongs to single individuals
ma	Membership Attack
nma	Non-membership Attack
sva	Sensitive Variance Attack
csa	Categorical Similarity Attack
QS-Loose Linkability	QT & ST linked one-to-many correspondence

BK = {QT, ST, BID, any publicly available information}, where BK is the background knowledge of an adversary.

Membership attack (ma): The privacy breach due to the identification of a particular sensitive value that belongs to an individual i can be linked with a specific group of QI attributes due to membership knowledge (mk) is called a membership attack.

Non-membership attack (nma): The privacy breach due to the identification of a particular sensitive value that belongs to an individual i can not be linked with a specific group of QI attributes due to non-membership knowledge (nmk) is called a non-membership attack.

Sensitive variance attack (sva): The low variability of SA values in an EC from different SA categories in a category table is called a sensitive variance attack.

Categorical similarity attack (csa): The SAs in an EC is obtained from a single category of the category table, which may narrow the adversary's knowledge to attack, called categorical similarity attack.

Sensitive vertical attack (sVer): The correlation and generalization of SA values vertically from different hierarchical levels, to isolate a SA value for re-identification of an individual, with the help of background knowledge called sensitive vertical attack.

QS-loose linkability: The proposed QT and ST are loosely (*i.e.,* independently) linked through one-to-many correspondence, for improved privacy and utility, instead of one-to-one tight correspondence is called loose likability.

Transformation (*Gong et al., 2017*): The records of same individuals have the same QID values in the dataset. The dataset can be transformed by merging all the same individual's QID values to a single set. In the transformed dataset each individual has only one record consisting of his/her QID and SAs.

Angelization (*Kanwal et al., 2019*; *Xiao & Tao, 2006*): The sensitive partitioning $A = \{A_1, A_2, \cdots, A_n\}$, and the quasi partitioning $B = \{B_1, B_2, \cdots, B_m\}$ of the microdata Table T, an Angelization of Table T produces two tables: ST and QT, such that, ST consists of Bucket-ID and SAs, where SAs represents the Sensitive Attribute column of Table T. QT contains all QI attributes with QID belonging to Table T.

High-Level Petri Nets (HLPN) (*Malik, Khan & Srinivasan, 2013*): A graphical and mathematical representation for examining the information control. It consists of 7-tuples; $N = (P, T, F, \phi, R_n, L, M_0)$. The static semantics are shown using L, ϕ and R_n whereas F, P, and T provide the dynamic structure. The P is the set of all places, where a single place is represented by a cycle. T is the set of all transitions (*i.e.,* rectangular boxes in HLPN), where transitions show the changes encounter in the system. The relation between P and T is such that $P \cap T = \phi$, $P \cup T \neq \phi$. The rules for these transitions are represented by R_n. F represents the information flow such that $F \subseteq (P \times T) \cup (T \cup F)$. The data types are mapped to the places P through ϕ, L refers to a label on F, and M_0 represents the initial marking.

HLPN ANALYSIS OF PREVIOUS MODELS

The formal modeling of (k, l)-diversity and θ-Sensitive k-Anonymity are performed here to reveal the way how an adversary can perform an attack.

(k, l)-diversity

The formal modeling and analysis reveals the way how an adversary can perform a sensitive vertical attack on the 1:M dataset as shown in Fig. 3. The working of (k, l)-diversity given in *Kanwal et al. (2019)* from Rule (1) to (12), where data is taken from data owner and data publisher anonymized it. The types are shown in Table 6 and data placed with description in Table 7. The black rectangular boxes show transition arrows showing the flow, and circles represent places or sub-part of the system. The data owner, data publisher, and the adversary are the entities.

The first transition shows input taken from the data owner of 1:M data after taking, anonymizing the data, and publishing it. After publishing that data adversary can perform

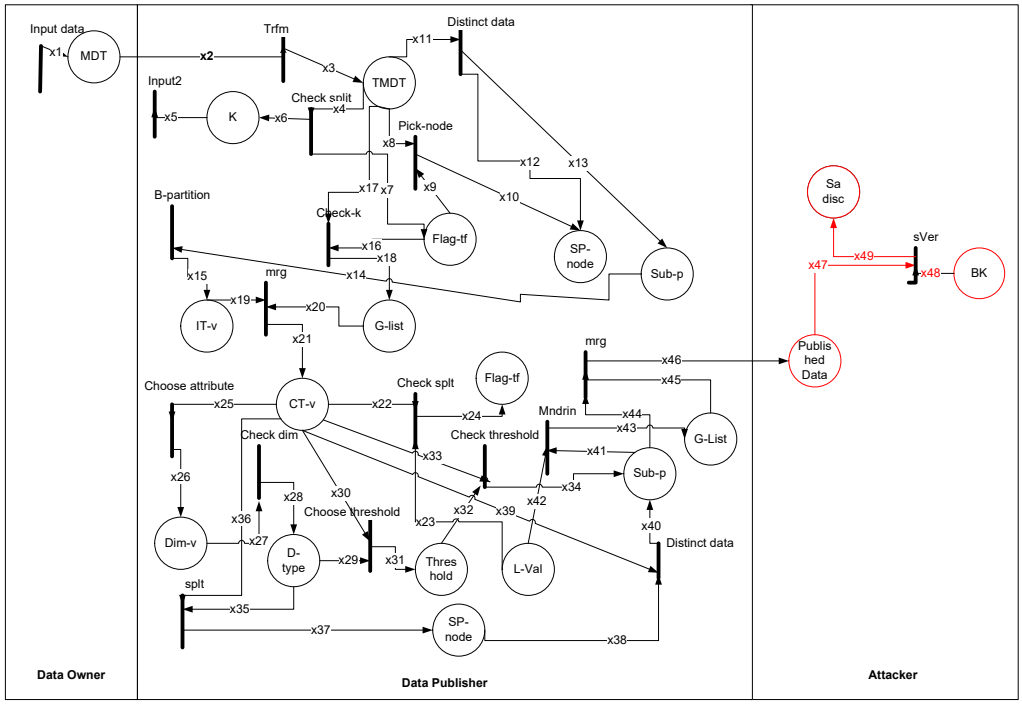

Figure 3 **HLPN for (k, l)-diversity.**

Full-size DOI: 10.7717/peerjcs.1255/fig-3

Table 6 **Types used in HLPN for (k, l)-diversity.**

Symbols	Description
PID	Patient ID
TID	Define level of k-anonymity
Contf	True or False
QI	Qausi idetifier
Int	Define QI using integer
Chr	Define QI using char
GSA	Generalized Sensitive Attribute
TSA	Transformed sensitive attribute
PID_{qi}	Reveal identity of patient using qausi attributes.

an attack on it. The anonymization process starts from transformation and then checks the individual record, distinct at least k-1. After this in each subpartition, the GSA record is balanced and after that stored in IT-v. The G-list consists of records less than k. The CT has contains records of IT-v and G-list. Then the category of numeric and categorical is checked by choose attribute and check dim, after this update values according to the threshold. Finally, Mndrn is used for generalization and checking the splits of SA subpartition, if it can not then places the records into G-list.

The sVer attack performed on (k, l)-diversity because of partially generalization of SA values. It can be correlated with any generalized categorical SA value. Since all the

Table 7 Mapping of data types in HLPN for (k, l)-diversity.

Symbols	Description
ϕ(MDT)	$\mathbb{P}$ (QI $\times$SA $\times$PID)
ϕ(TMDT)	$\mathbb{P}$ (QI $\times$SA $\times$TID)
ϕ(Flag-tf)	$\mathbb{P}$ (Condtf)
ϕ(G-list)	$\mathbb{P}$ (QI $\times$SA $\times$TID)
ϕ(k)	$\mathbb{P}$(k)
ϕ(L)	$\mathbb{P}$(l)
ϕ(SP-node)	$\mathbb{P}$ (GSA)
ϕ(Sub-p)	$\mathbb{P}$ (QI $\times$GSA $\times$TID)
ϕ (IT-v)	$\mathbb{P}$ (QI $\times$ITSA $\times$TID)
ϕ(CT-v)	$\mathbb{P}$ (QI $\times$ITSA $\times$TID)
ϕ(Dim-v)	$\mathbb{P}$ (Int$\times$ chr)
ϕ(D-type)	$\mathbb{P}$ (Int$\times$ chr)
ϕ(Thres hold)	$\mathbb{P}$ (Thr-v)
ϕ(Published Data)	$\mathbb{P}$ (QI $\times$ SA)
ϕ(BK)	$\mathbb{P}$ (PID $\times$QI $\times$SA)
ϕ(Sa disc)	$\mathbb{P}$ (QI $\times$SA $\times$PID)

generalized categorical SA values are on a different level in the hierarchy, SA values on a different level can be vertically correlated to breach the privacy of an individual in Rule 3 as.

$$\mathbf{R}(\mathbf{sVer}) := \forall i47\epsilon, x47, \forall i48\epsilon x48, \forall i49\epsilon x49 | sVer - atk(i47[2], i48[2]) \rightarrow (i49[1] = i2[2] \wedge i49[2]) = i2[3]. \quad (3)$$

In Rule 3 an attacker can be attacked due to the lowest common cut on transaction generalization hierarchy, leaving the single SA from different sub-set unprotected and vulnerable which leads toward correlation of the background information with published data that ultimately caused an attack on sensitive attributes known as sVer.

θ-Sensitive k-Anonymity

The formal modeling analysis reveals the way how an adversary can perform a sensitive variance attack on 1:M data because of inapplicability of θ-Sensitive k-Anonymity for 1:M microdata.

The working of theta given in *Khan et al. (2020a)* from Rule (1) to (8), where data is taken from data owner and data publisher anonymized it as shown in Fig. 4. The types are shown in Table 8 and data placed with description in Table 9. The black rectangular boxes show transition arrows showing the flow, and circles represent places or sub-part of the system. The data owner, data publisher, and the adversary are the entities.

The first transition shows input is taken from the data owner of 1:M data after taking, anonymizing the data, and publishing. After publishing that data adversary can perform an attack on it. The anonymization process starts from taking input and checking k value. After it calculates threshold using var() function, and if needed variance value adjust using adj(), swap using swap() or adding noise using Add ns(). For 1:M dataset, SA values are

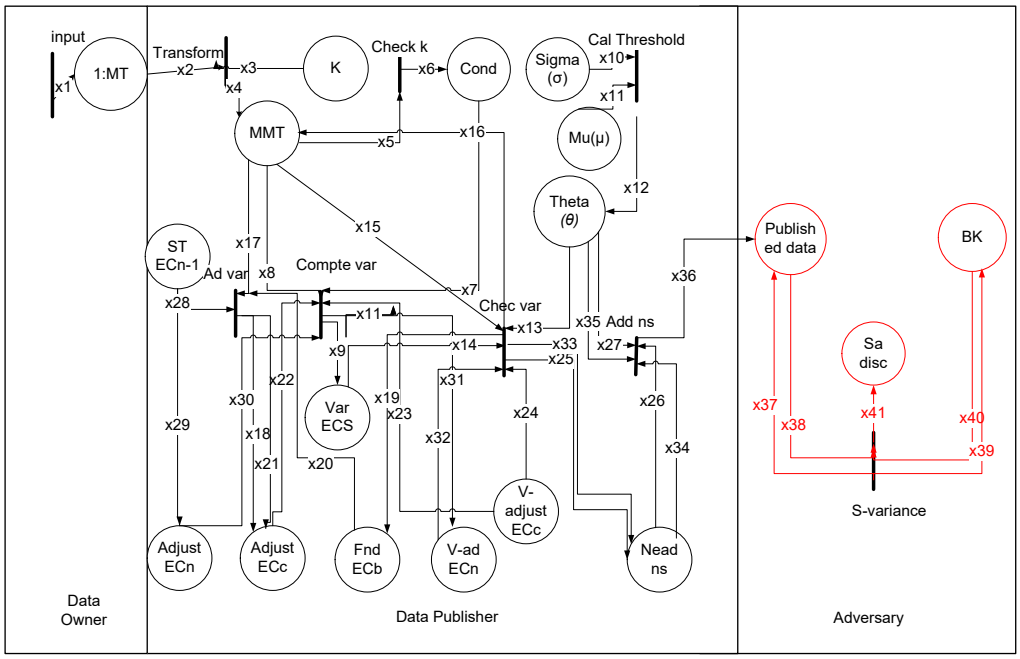

Figure 4 **HLPN for θ-Sensitive k-Anonymity.**
Full-size DOI: 10.7717/peerjcs.1255/fig-4

Table 8 **Types used in HLPN for θ-Sensitive k-Anonymity.**

Symbols	Description
M	EC size
Condition	True or False
σ	The notation of sigma
μ	The notation of observation value
θ	The notation of threshold value
fnd EC_b	Found the EC of b
Adjust EC_c	The EC of c adjust
Adjust EC_n	The EC of n adjust
Var ECS	Different ECS varaince.
V-ad EC_n	The variance of EC adjust for class n
V-adjust EC_c	The variance of EC adjust for class c

more than one in a single tuple in that way θ-Sensitive k-Anonymity approach can not be directly applied to implement diversity in an EC. Because the variance-based θ threshold calculation to obtain a diverse EC is not applicable for 1:M microdata. So a sensitive variance attack can be performed on the published data using any background knowledge. In Rule 4 as:

$$\mathbf{R(SVA)} := \forall i38\epsilon, x38, \forall i40\epsilon x40, \forall i41\epsilon x41, \forall i2\epsilon x2 | sva - att(i38[2], i40[2]) \rightarrow i43[1] = i2[1] \wedge i41[2] = i2[3] \quad (4)$$

Table 9 Mapping of data types in HLPN for θ-Sensitive k-Anonymity.

Symbols	Description
ϕ(1:MT)	$\mathbb{P}$ ($A^{id} \times A^{qi} \times A^{sa}$
ϕ(MMT)	$\mathbb{P}$ ($EC_C \times EC_b \times EC_n \times k$)
ϕ(k)	$\mathbb{P}$ (k)
ϕ(cond)	$\mathbb{P}$(Condition)
ϕ(sigma)	$\mathbb{P}$ (σ)
ϕ(mu)	$\mathbb{P}$ (μ)
ϕ(theta)	$\mathbb{P}$ (θ)
ϕ(Fnd EC_b)	$\mathbb{P}$ (EC_b)
ϕ(Var ECS)	$\mathbb{P}$ ($VEC_c \times VEC_b \times VEC_n$)
ϕ(Adjust EC_c)	$\mathbb{P}$(EC_c)
ϕ(Adjust EC_n)	$\mathbb{P}$ (EC_n)
ϕ(ST EC_{n-1}	$\mathbb{P}$ (EC_{n-1})
ϕV-adjust EC_c	$\mathbb{P}$ ($V_{EC\,c}$)
ϕV-adjust EC_n	$\mathbb{P}$ ($V_{EC\,n}$)
ϕNead ns	$\mathbb{P}$ ($VEC_c \times A^{id} \times A^{qi} \times A^{sa}$)
ϕP data	$\mathbb{P}$ ($A^{qi} \times A^{sa}$)
ϕBK	$\mathbb{P}$ ($A^{id} \times A^{qi}$)
ϕSa disc	$\mathbb{P}$ ($A_i^{id} \times A_i^{sa} \times A_i^{qi}$)

The EC produced, can not prevent the sva in definition 3, and csa in definition 4, because of inappliabilty of θ-Sensitive k-Anonymity for 1:M dataset, therefore the attack can be performed on 1:M dataset.

PROPOSED (θ^*, K)-UTILITY

The purpose of anonymization should not be singleton to either privacy or utility. And at the same time the proposed approach must be strong enough to prevent any possible attack and also provide quality of data. Therefore, the purposed (θ^*, k)-utility algorithm in this article not only anonymizes 1:M data to prevent possible attacks, *e.g.*, *ma*, *nma*, *sva*, *csa*, and *sVer*, but also provide improved quality data.

In our proposed approach we apply full generalization of SA, that can bee seen in the transformation of Table 4 into Table 10. Further, we apply partitioning of the QI and SA, and use QS-loose linkability for one-to-many correspondence to prevent any possible privacy leakage. To prevent the sVer attack, identified in (k, l)-diversity (*Gong et al., 2017*), the SA are placed in ECs using the θ-Sensitive k-Anonymity (*Khan et al., 2020a*) approach. However, since the anonymized data is 1:M, and the θ-Sensitive k-Anonymity (*Khan et al., 2020a*) cannot be directly applied because of the variance calculation for each EC. Therefore, the leaf-nodes of SA in Fig. 2 are generalized based on CtgT Table 3 and applying the θ-Sensitive k-Anonymity (*Khan et al., 2020*) to implement required diversity in each EC. The θ-Sensitive k-Anonymity (*Khan et al., 2020*) approach is a simple numerical measure of privacy strength which ensures a strong privacy implementation for each EC and hence for the complete dataset.

Table 10 Transformed Microdata T.

Patient Record ID	Age	Zipcode	Gender	Symptoms
p1	40	2139	Female	Mild
p2	45	2545	Male	Severe
p3	45	2238	Female	High
p4	38	2843	Male	Severe
p5	42	2341	Female	Low
p6	40	2548	Male	Low
p7	45	2544	Female	A-symptomatic
p8	32	2538	Male	Low

Proposed (θ^*, k)-utility: The sensitive partitioning SA $=\{A_1^{sa}, A_2^{sa}, A_3^{sa}, \cdots, A_n^{sa}\}$, and the quasi partitioning QI $=\{A_1^{qi}, A_2^{qi}, A_3^{qi}, \cdots, A_n^{qi}\}$, of the transformed 1:M microdata T into 1:1 microdata linked through QS-Loose linkability, that produces two tables: Sensitive Table (ST) and Quasi Table (QT). The ST consists of SA and BID, where the QT consists of age, zipcode, gender and BID. Below Eq. (5) depicts the proposed approach.

$$\begin{aligned} &\text{iff}|\forall \tau_i \epsilon \mathscr{T} : \{A_n^{sa} \leftarrow \text{Count}(\text{Dist}(\text{A}_i^{sa})) \leq \theta\}| \geq 2k \wedge \\ &(\forall \tau_i : \{A_i^{sa} \cdot BID\} \epsilon \text{sbt} \wedge \{A_i^{si} \cdot BID\} \epsilon \text{genData}) \end{aligned} \tag{5}$$

where τ_i represents the tuples from the complete dataset $\mathscr{T}$, having the maximum SA values belonging to CtgT in a transformed 1:1 record shape. So, in the first half of the equation for creating the ST, the proposed approach will execute for checking the θ condition from the θ-Sensitive k-Anonymity approach if the total number of tuples are greater than the user input (*i.e.,* k size, read proposed Algorithm 1 at line 21). The second part of the equation finalizes the *sbt* and *genData* (*i.e.,* the ST and QT respectively), and which are the tables obtained through the proposed (θ^*, k)-utility algorithm.

The (θ^*, k)-utility Algorithm

The working of the proposed (θ^*, k)-utility Algorithm 1 is partitioned into three major parts; transformation (lines 3-18), sensitive buckets creation (line 20-27), and quasi generalized buckets creation (lines 29-35). Initially, the 1:M data in T is in its original form, and the sensitive buckets (sbt) and the quasi generalized data (genData) are taken as empty sets. The algorithms begin by computing distinct SAs in T (lines 3-5), which are further categorized into five different sensitive categories to create category CtgT at line 7 (Table 3).

Algorithm 1 (θ^*, k)-utility

Require:
T: 1:M Microdata Table;
k: k-anonymity;
Γ: T populated with CtgT ;
τ: Individual tuple from transform Table ($\mathscr{T}$);
$\hbar$: Individual history tuple from history table $\mathbb{H}$;
Ensure:
QT: Quasi Table :-genData;
ST: Sensitive Table :-sbt;

1: sbt={};
2: genData={};
3: **for** all $t_i^{sa} \cdots t_n^{sa}$ in T **do**
4: D_n^{sa} := Compute(Distinct(SA value))
5: **end for**
6: **for** all $d_i^{sa} \cdots d_n^{sa}$ **do**
7: CtgT := Categorize D_n^{sa} into five categories
8: **end for**
9: **for** all $t_i^{sa} \cdots t_n^{sa}$ **do**
10: Γ:= T $\leftrightarrow CtgT^{cat}$
11: **end for**
12: **for** all $r_i \cdots r_n$ **do**
13: **for** all $t_i \cdots t_n$ **do**
14: $\tau_i := max(t_i^{sa})_{\forall t_i^{sa} \epsilon\, CtgT^{cat}}$
15: $\hbar_i := \neg max(t_i^{sa})_{\forall t_i^{sa} \epsilon\, CtgT^{cat}}$
16: **end for**
17: **end for**
18: $\mathscr{T} := \sum_{i=1}^{n} \tau_i$
19: $\mathbb{H}_i := \sum_{i=1}^{n} \hbar i$
20: **while** $\mathscr{T} \neq \{\}$ **do**
21: **if** $\mathscr{T} \leq 2k$ **then**
22: $sbt_k :=$ $\mathscr{T}$
23: sbt := sbt $\cup sbt_k$
24: **else**
25: Apply θ-Sensitive k-Anonymity Khan, Razaullah and Tao, Xiaofeng and Anjum, Adeel and Kanwal, Tehsin and Malik, Saif Ur Rehman and Khan, Abid and Rehman, Waheed Ur and Maple, Carsten (2020)
26: **end if**
27: **end while**
28: N := $|\mathscr{T}^{qi}|$ // BID is obtained from bk^{sa}
29: **while** $N \neq \{\}$ **do**
30: **if** $N \leq 2k$ **then**
31: genData := N
32: **else**
33: genData := genData $\cup$ gen(N) // Linked via BID
34: **end if**
35: **end while**
36: **return** sbt
37: **return** genData

The CtgT will be used as a reference table while creating sbt. Lines 9-11, populate the original 1:M Table T by assigning the categorical SA values (*i.e., Low*, *Mild*, *High*, *Severe*, *A-symptomatic*) to the actual SA (*i.e.,* symptoms attribute), shown in Table 4. The *for* loop (lines 12-17), creates transform table $\mathscr{T}$; Table 10 at line 14 (*i.e.,* see definition 7). Table 10 has 1:1 microdata. Lines 12 and 13 checks the number of tuples (t_i) that belongs to a single individual record (r_i), i.e. 1:M data. Line 14 creates the transformed tuples (τ_i) by selecting high weighted categorical sensitive attribute values. The sensitive vertical attack (sVer) in

Table 11 History table.

Patient Record ID	Tuple ID	Name	Age	Zipcode	Gender	Symptoms
p1	t1	Susan	40	2139	Female	Low
	t2	Susan	40	2139	Female	Mild
p2	t4	Ronald	45	2545	Male	High
p3	t6	Keran	45	2238	Female	Mild
p4	t8	Heather	38	2843	Male	High
p5	t10	Cytnthia	42	2341	Female	Low
	t11	Cytnthia	42	234a1	Female	Low
p6	t14	Peter	40	2548	Male	Low

definition 5 is prevented at this step of the algorithm. By creating the transformed Table 10 from the original Table 1, the leaf-nodes (*i.e.,* in Fig. 2) cannot be correlated with any generalized categorical SA value. Since all the generalized categorical SA values are on the same level in the hierarchy, SA values on a single level cannot be vertically correlated to breach the privacy of an individual. The remaining categorical sensitive values are stored in a history table $\mathbb{H}$ (Table 11) at line 19, to avoid any wastage of data.

Next, the algorithm will create sensitive buckets from the transformed data $\mathscr{T}$. The *while* loop processes all the tuples in $\mathscr{T}$ to create k (*i.e.,* user input) size sensitive buckets (sbt). If the tuples to anonymize in $\mathscr{T}$ are less than 2 k the algorithm will create final sbt, otherwise it will process sensitive part of all tuples using the θ-Sensitive k-Anonymity algorithm (*Khan et al., 2020*), to create more diverse Equivalence Classes (ECs) for ST (*i.e.,* Table 12). Since the sbt obtained through the θ-Sensitive k-Anonymity algorithm (*Khan et al., 2020*), the EC produced can prevent the *sva* (*i.e.,* Definition 3), and *csa* (*i.e.,* definition 4). The prevention of data from *sva* and *csa* have already been proved in the algorithm of θ-Sensitive k-Anonymity (*Khan et al., 2020*).

The last part of the algorithm (line 29-35) creates a generalized QT *i.e.,* Table 12A. The QIs are anonymized at lines 31 and 33 in such a way that the adversary's confidence about the presence of an individual (in definition 1: ma) or confidence over the absence (*i.e.,* Definition 2-nma) is prevented. The obtained k-anonymized quasi buckets (kb) are linked through the Bucket ID (BID) with the sbt using the one-to-many correspondence (QS-Loose linkability) between the two sub-tables, *i.e.,* Tables 12A and 12B. In Table 12A, the Patient Record ID column is not part of the published table. Finally, the algorithm returns the genData in the form of QT, and sbt in the form of ST, linked through BIDs. The tables obtained from the proposed Algorithm 1 are shown in Tables 12A and 12B. The one-to-many correspondence between QT and ST is the loose linkage (in definition 6) between a single sbt in ST with more than one tuples in different ECs in QT. The EC4 in Table 12A adds a noise tuple (*i.e.,* n1) correspondent to the already added noise SA value in Table 12B because of the θ requirements in θ-Sensitive k-Anonymity algorithm (*Khan et al., 2020*). The beauty of the QS-loose linkability is the improved utility of the data. Because it allows the least distance QI values to create an EC that can be linked with more than one sbt in ST. Another beauty is the improved privacy implementation. Because the

Table 12 Anonymized data obtained via proposed (θ^*, k)-utility algorithm.

(a) Quasi table (QT)

Patient Record ID	Age	Zipcode	Gender	BID
p1	(40-42) [40]	(2139-2341) [2319]	Female	1
p5	(40-42) [42]	(2139-2341) [2341]	Female	3
p2	(40-45) [45]	(2545-2548) [2545]	Male	1
p6	(40-45) [40]	(2545-2548) [2548]	Male	4
p3	(45-45) [45]	(2238-2544) [2238]	Female	2
p7	(45-45) [45]	(2238-2544) [2544]	Female	3
p4	(32-38) [38]	(2538-2843) [2843]	Male	2
p8	(32-38) [32]	(2538-2843) [2538]	Male	4
n1	(32-38) [32]	(2538-2843) [2538]	Male	4

(b) Sensitive table (ST)

Symptoms	BID
Mild	1
Severe	
High	2
Severe	
Low	3
A-symptomatic	
Low	4
Low	
Mild	

adversary's confidence to uniquely identify a tuple that belongs to an individual, is reduced by linking a single sbt in ST with more than one k-anonymized ECs in QTs.

HLPN analysis of (θ^*, k)-utility algorithm

The different attacks discussed in Section 5 and 6 are mitigated through the proposed (θ^*, k)-utility algorithm.

The data owner, data publisher, adversary are used to model HLPN for the (θ^*, k)-utility algorithm in Fig. 5. The types showed in Table 13 and data places with description in Table 14. For (θ^*, k)-utility algorithm initially find distinct SAs values. The Rule 6, is used to compute distinct SAs sensitive attributes. In Rule 6 as:

$$\mathbf{R}(\mathbf{Compute}(\mathbf{Dist}(\mathbf{SA}))) := \forall i2 \epsilon x2, i3 \epsilon x3 | i3[1] := \text{Dist(SA)}\ (i2[1]) \wedge x3^{'} := x3 \cup \{i3[1]\} \quad (6)$$

After this by using Rule 7, categorized symptoms into five different sensitive categories to create category Table 3. In Rule 7 as:

$$\mathbf{R}(\mathbf{Categorize}) := \forall i4 \epsilon x4, i5 \epsilon x5 | i5[1] := i4[1] \wedge i5[2] := i5[2] \wedge x5^{'} := x5 \cup \{i5[1], i5[2]\} \quad (7)$$

The Rule 8, populate the Table 1 by assigning the categorical SA values (*i.e.,* Low, Mild, High, Severe, A-symptomatic) to the actual SA (*i.e.,* symptoms attribute) in Table 4. In Rule 8 as:

$$\mathbf{R}(\mathbf{Populate}) := \forall i6 \epsilon x6, i7 \epsilon x7 | i7[1] := \text{Populate i6}[1] \wedge x7^{'} := x7 \cup \{i7[1]\} \quad (8)$$

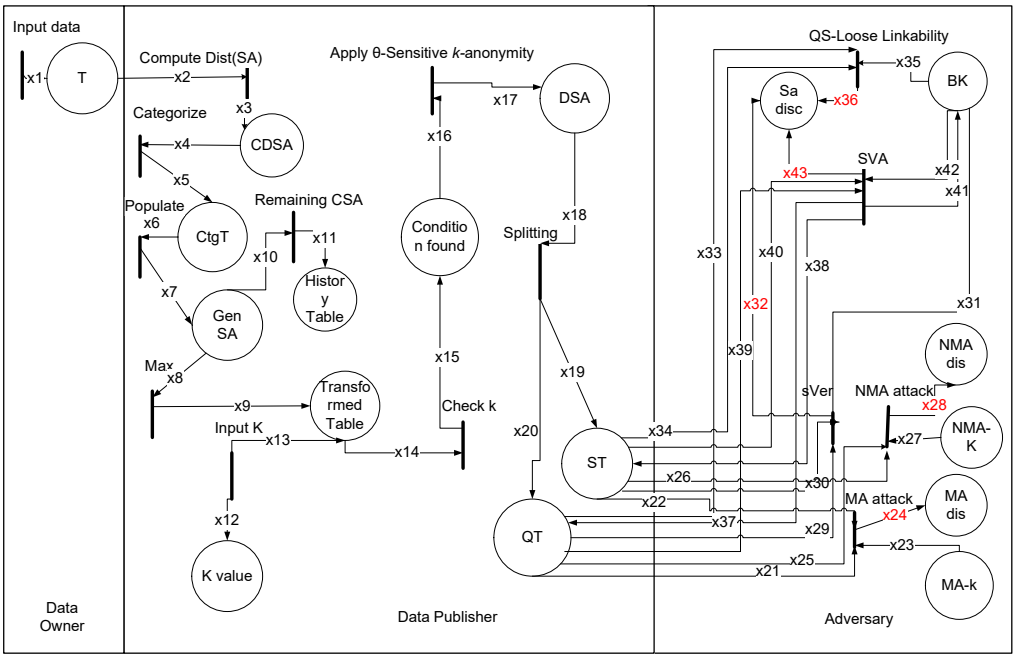

Figure 5 **HLPN for (θ^*, k)-utility algorithm.**

Full-size DOI: 10.7717/peerjcs.1255/fig-5

Table 13 **Types used in HLPN for (θ^*, k)-utility.**

Symbols	Description
DSA	Distinct set of sensitive attributes
SAC	Sensitive attributes consist of 5 categories
GSA	Generalized SA
RRT	Remaining Repeated Tuples
k	Input for ECs
CF	Condition for k size
QS	Set of implemented θ ECs
A_{nm}^{sa}	Non-membership knowledge for sensitive values
A_{ma}^{sa}	Membership knowledge for sensitive values
τ^i	Transform into 1:1 tuples

Rule 9 is used for selecting a high weighted categorical attributes sensitive value to transformed in Table 10. In the transformation process ignored records stored in History Table 11, it consists of remaining categorical sensitive values, that are stored in a History Table 11 to avoid any wastage of data in Rule 10. The 2 k input take using Rule 11. The k value checked using Rule 12. The threshold value calculated using Rule 13, it will process sensitive part of all tuples using the θ-Sensitive k-Anonymity algorithm (*Khan et al., 2020*), to create more diverse Equivalence Classes (ECs) for Sensitive Table (ST) Table 12B.

$$\mathbf{R}(\mathbf{Max}) := \forall i8 \epsilon x8, i9 \epsilon x9 | i9[1] := \max(i8[1]) \wedge x9^{'} := x9 \cup \{i9[1]\} \quad (9)$$

Table 14 Mapping of data types in (θ^*, k)-utility.

Symbols	Description
ϕ(T)	$\mathbb{P}$ ($A^{id} \times A^{qi} \times A^{sa}$)
ϕ(CDSA)	$\mathbb{P}$ (DSA)
ϕ(CtgT)	$\mathbb{P}$ (SAC)
ϕ(GenSA)	$\mathbb{P}$ (GSA)
ϕ(History Table)	$\mathbb{P}$(RRT)
ϕ(Transformed Table)	$\mathbb{P}(\tau^i)$
ϕ(K-value)	$\mathbb{P}$ (k)
ϕ(Condition found)	$\mathbb{P}$ (CF)
ϕ(DSA)	$\mathbb{P}$ (QS)
ϕ(QT)	$\mathbb{P}$ ($A^{qi} \times$ BID)
ϕ(ST)	$\mathbb{P}$ ($A^{sa} \times$ BID)
ϕ(BK)	$\mathbb{P}$($A^{id} \times A^{qi} \times A^{sa}$)
ϕ(SA disc)	$\mathbb{P}$ ($A_i^{id} \times A_i^{qi} \times A_i^{sa}$)
ϕ(MA disc)	$\mathbb{P}$ ($A_i^{id} \times A_i^{qi} \times A_{ma}^{sa}$)
ϕ(NMA disc)	$\mathbb{P}$ ($A_i^{id} \times A_i^{qi} \times A_{nm}^{sa}$)

$$\mathbf{R}(\textbf{Remaining CSA}) := \forall i10 \epsilon x10, i11 \epsilon x11 | i11[1] := \neg \max\,(i10[1]) \wedge x11^{'} := x11 \cup \{11[1]\} \tag{10}$$

$$\mathbf{R}(\textbf{Input K}) := \forall i12 \epsilon x12, i13 \epsilon x13 | i13[1] := \text{input}\,(i12[1]) \wedge x13^{'} := x13 \cup \{13[1]\} \tag{11}$$

$$\mathbf{R}(\textbf{Check k}) := \forall i14 \epsilon x14, i15 \epsilon x15 | i15[1] := \text{check}\,(i14[1]) \wedge x15^{'} := x15 \cup \{15[1]\} \tag{12}$$

$$\mathbf{R}(\textbf{Apply } \theta\textbf{-Sensitive k-Anonymity}) := \forall i16 \epsilon x16, i17 \epsilon x17 | i17[1] := \theta\text{-Sensitive k-Anonymity}(i16[1]) \wedge x17^{'} := x17 \cup \{17[1]\} \tag{13}$$

The splitting is performed for QA and SA attributes, and both are linked with one-to-many correspondence using Rule 14. The obtained k-anonymized quasi buckets are linked through the Bucket ID (BID) with the sbt using the one-to-many correspondence between the two tables. Finally, the algorithm returns the Table 12A QT and Table 12B ST linked through BIDs. In Rule 14 as:

$$\begin{aligned}\mathbf{R}(\textbf{Splitting}) := \forall i18 \epsilon x18, i19 \epsilon x19, i20 \epsilon x20 | \\ i19[1] := Split(\{i18[1]\}) \wedge i19[2] := \\ BID\{i18[1]\} \wedge x19^{'} := x19 \cup \{i19[1], i19[2]\} \\ i20[2] := Split(\{i18[1]\}) \wedge i20[2] := \\ BID\{i18[1]\} \wedge x20^{'} := x20 \cup \{i20[1], i20[2]\}\end{aligned} \tag{14}$$

$$\mathbf{R}(\mathbf{MA\ Attack}) := \forall i21\epsilon x21, \forall i22\epsilon x22, \forall i23\epsilon x23, \forall i24\epsilon x24 | MADis(i21[1], i22[2]) \rightarrow i21[1]i22[1] \cup i23[2] \neq i2[2] \wedge i2[3](i24[2] \cup i24[3]) = \varnothing \quad (15)$$

$$\mathbf{R}(\mathbf{NMA\ Attack}) := \forall i25\epsilon x25, \forall i26\epsilon x26, \forall i27\epsilon x27, \forall i28\epsilon x28 - NMADis(i25[1], i27[2]) \rightarrow i28[2] \wedge NMADis(i26[1], i27[2]) = \varnothing \quad (16)$$

$$\mathbf{R}(\mathbf{SVer\ Attack}) := \forall i29\epsilon x29, \forall i30\epsilon x30, \forall i31\epsilon x31, \forall i32\epsilon x32 | SaDisc(i29[1], i30[1]) \rightarrow (\{i29[1], i130[1]\} \cup \{i31[2]\}) := i2[2] \wedge i2[3](i32[2] \cup i32[3]) = \varnothing \quad (17)$$

$$\mathbf{R}(\mathbf{QS-Loose\ Linkability}) := \forall i33\epsilon x34, \forall i34\epsilon x34, \forall i35\epsilon x35, \forall i36\epsilon x36 | (i33[1], i34[1]) \rightarrow (\{i33[1], i134[1]\} \cup \{i35[2]\}) := i2[2] \wedge i2[3](i36[2] \cup i36[3]) = \varnothing \quad (18)$$

$$\mathbf{R}(\mathbf{SVA}) := \forall i39\epsilon, x39, \forall i40\epsilon x40, \forall i42\epsilon x42, \forall i43\epsilon x43 | SaDisc(i39[2], i40[2], i42[2]) \neq (i2[1] \cup i2[2] \cup i2[3])(i43[2] \cup i43[3]) = \varnothing \quad (19)$$

In Rules 15 and 16, the adversary's confidence about the presence of an individual, or confidence over the absence is prevented. The sensitive vertical attack (sVer) prevented by creating the Table 4 from the Table 1, so SA values on a single level cannot be vertically correlated to breach the privacy of an individual in Rule 17. The adversary's confidence to uniquely identify a tuple which belongs to an individual, is reduced by linking a single sbt in ST with more than one k-anonymized EC in QTs in Rule 18. The diverse EC produced, can prevent the sva and csa. The prevention of data from sva and csa in Rule 19. The (θ^*, k)-utility algorithm protects from above mentioned attacks, results in form of a null value as shown in Rules 15, 16, 17, 18, and 19.

EXPERIMENTAL ANALYSIS

In this section, we analyse the comparative results of our proposed (θ^*, k)-utility technique for 1:M microdata, in terms of utility, privacy, and computational efficiency.

The anonymized data quality and the execution time is measured to compare proposed (θ^*, k)-utility algorithm, with θ-Sensitive k-Anonymity (*Khan et al., 2020*), (k, l)-diversity (*Gong et al., 2017*), and (p, l)-angelization (*Kanwal et al., 2019*), algorithms.

Experimental setup

All the experiments are performed on a machine having Windows 10 operating system with Core i7 processor and 8GB RAM. The proposed algorithm is implemented in Python 3.9 language. We used the modified 'Adult' dataset, which is publically accessible from the repository of UC Irvine machine learning.[1] [,2]

[1] DOI: 10.5281/zenodo.7214275, also in original can be seen on: https://archive.ics.uci.edu/ml/datasets

[2] Some examples of modified datasets were used in other research as a reference. https://archive.ics.uci.edu/ml/datasets, https://datahub.io/machine-learning/adult#readme, https://www.researchgate.net/figure/Examples-of-generated-counterfactuals-on-the-modified-Adult-dataset-Example-Based-CF-and_tbl2_337830079, https://openreview.net/forum?id=bYi_2708mKK

In the modified Adult dataset the *age*, *zipcode*, and *gender* are considered as QIs, while the *symptoms* attribute is considered as the SA.

The anonymized data obtained from the proposed and the base algorithms are analyzed for utility using normalized certainty penalty (NCP) (*Anjum et al., 2018*) and query accuracy (*Anjum et al., 2018*), for privacy using the average number of vulnerable records, and the computational efficiency is analyzed with the average execution time of all the algorithms.

Utility loss

The utility loss of the anonymized data is measured using the following techniques.

Normalized certainty penalty

Normalized certainty penalty (NCP) is one of the techniques which measures the utility loss caused by data anonymization. We measure the utility loss caused by the QIs. High penalties indicate high utility loss and vice versa.

Let T $=\{q_1,q_2, \ldots,q_m\}$ are QI. The utility loss for a single QI attribute is shown in Eq. (20) as.

$$NCP_{q_i}(t) = \frac{x_i - y_i}{|Qi|} \tag{20}$$

where $y_i \leq z_i \leq x_i$, and z_i is the actual QI value from T, and $|Qi$—is the domain range on QI_i, *i.e.*, $max\{t.QI_i\} - min\{t.QI_i\}$. The total weighted certainty penalty for the whole table is the sum of all attributes in a tuple and then adding NCP obtained from all tuples, as in Eq. (21).

$$NCP(T^*) = \sum_{t=T^*} \sum_{i=1}^{q} w_i \cdot NCP_{qi}(t) \tag{21}$$

where, NCP(t) $= \sum_{i=1}^{q} w_i \cdot NCP_{q_i}$ (t) represents penalty for a tuple, w_i are weights associated to attributes, and T^* is the final anonymized release.

Figure 6 shows the NCP for utility measurement on anonymized release. Figure 6 shows the comparative results of θ-Sensitive k-Anonymity, (k, l)-diversity, (p, l)-angelization and the proposed (θ^*, k)-utility algorithms, with varying k for the complete dataset. The increase in the graph values shows more utility loss. The higher value of k collectively for all algorithms results in higher utility loss because of the increased generalization range in each EC. For (k, l)-diversity, it is impossible to satisfy both k-anonymity and l-diversity constraints at the same time to achieve high privacy with minimum information loss, where the high value of the l-diversity is not recommended for high value of k. However, still the (k, l)-diversity results shows better utility than the θ-Sensitive k-Anonymity and the (p, l)-angelization algorithms, because the (p, l)-angelization algorithm only extends the (k, l)-diversity (which works only for 1:1 dataset) for the 1:M-MSA data without considering utility and privacy of the anonymized data. In (p, l)-angelization, increasing the diversity in for sensitive attributes in SAFBs greatly reduces the utility of the k-anonymous groups because of the one-to-one correspondence between the ST and QT. Through the one-to-one correspondence between the ST and QT, any change in ST

directly affect the QT for the same changes. So with respect to utility (p, l)-angelization is more worse than (k, l)-diversity. The multiplicative increase in utility loss for θ-Sensitive k-Anonymity is because of its straight forward privacy implementation for single SA without any contribution for utility improvement in the developed algorithm.

Our proposed (θ^*, k)-utility algorithm, independently generalizes the QI values to create less distance ECs for any size of k, which results in low utility loss as compared to its counterparts. Here, the k size in QT is not affected by any changes in ST. The proposed (θ^*, k)-utility algorithm in Fig. 6B depicts the NCP with varying size of data set and for a fixed value of k =4. The utility loss reduces with more and more 1:M records is because of the availability of more suitable QI values from the increased dataset to create smaller distance ECs, which reduces the loss in data utility. Our proposed (θ^*, k)-utility algorithm produces better results as compared to its counterparts, and depicts almost a constant data utility with any number of records. This is because of the separate publishing of the QT from the ST, which is enabled by the one-to-many correspondence approach.

Query accuracy

Query accuracy measures the utility loss between the original and anonymized release using an aggregate query, *e.g.* COUNT, AVG, SUM etc. Consider the following aggregate query in Eq. (22).

$$SELECT\ COUNT(*) \text{from T}^* \text{ where } \mathrm{A}_1^{qi} \epsilon \text{ Domain } (A_1^{qi}) \text{ AND} \cdots \\ \text{AND } \mathrm{A}_m^{qi} \epsilon\ Domain\ (A_m^{qi}) \tag{22}$$

Anonymized table T^* has m as a total A^{qis}, *i.e.*, $A_i^{qi}, A_i^{qi}, \ldots, A_q^{qi}$. The domain size *i.e.*, (A_i^{qi}) depends on query selectivity (θ) which indicates the expected number of tuples selection from an executed aggregate query. The tuple selectivity can be seen in in Eq. (23).

$$\theta = \frac{|t_q|}{|T|} \tag{23}$$

where $|T|$ is the total number of tuples in the dataset and $|t_q|$ indicates the number of tuples obtained from a query (Q). To measure the utility loss, the query error in Eq. (24) analyzes the error between the COUNT queries executed on the published and original dataset.

$$QueryError = |Count(anonymized) - Count(original)|/Count(original) \tag{24}$$

Query error is a common matrix to measure the utility of the anonymized release. We perform utility loss analysis between the θ-Sensitive k-Anonymity, (k, l)-diversity, (p, l)-angelization, and the proposed algorithm of (θ^*, k)-utility, by generating 1000 randomly queries and averaging their query error in Fig. 7.

Figure 7A shows the query error for varying k size. For all the comparative graphs, with increase in k size the query error increases, because high k means high distance ECs. Then the aggregate query results in more number tuples on anonymized data as compared to the original data. So the high comparative difference between the original and anonymized data results in an increased query error rate. The (p, l)-angelization do no focus on the utility of the data at all because even the separately created sensitive and quasi tables are

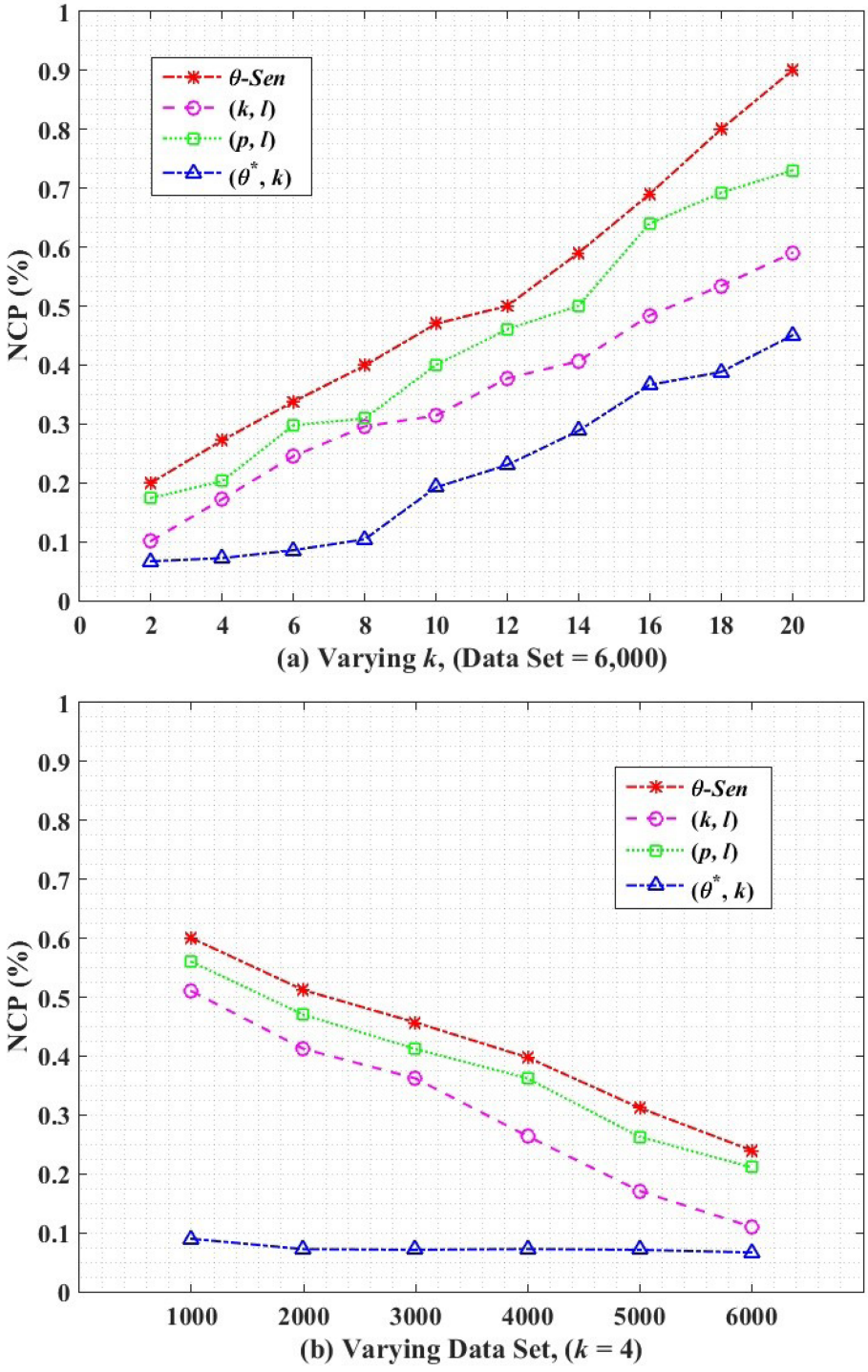

Figure 6 Normalized certainty penalty.

Full-size DOI: 10.7717/peerjcs.1255/fig-6

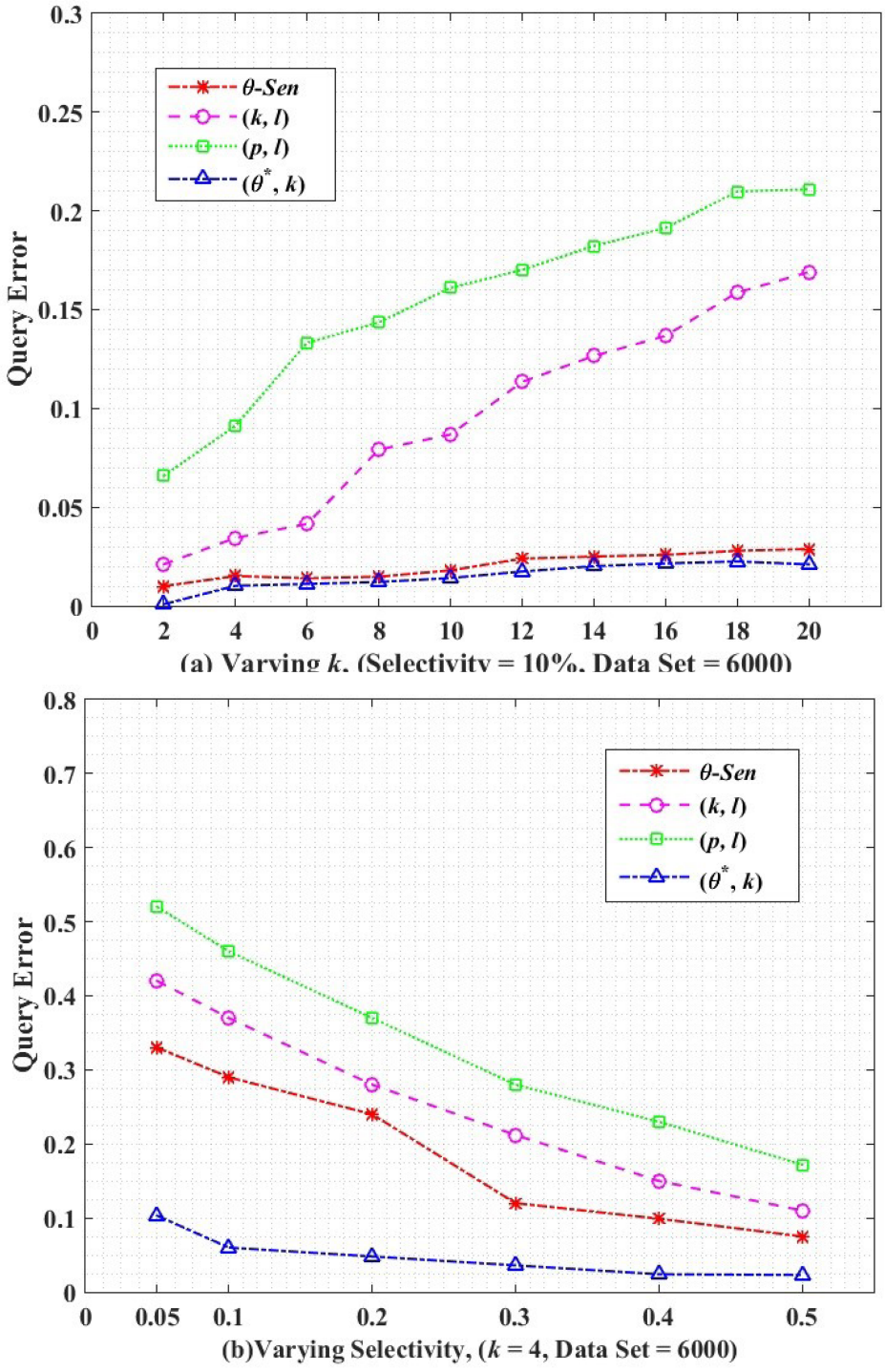

Figure 7 Query error.

Full-size DOI: 10.7717/peerjcs.1255/fig-7

not considered as actual separate tables because they are still directly connected through the one-to-one buckets in both tables. While in (k, l)-diversity although both QI and SA are generalized separately, even though, both are dependent on k, for measuring the utility of any k-anonymous group. The values obtained from the θ-Sensitive k-Anonymity indicates its better utility than the (p, l)-angelization and (k, l)-diversity is because of its local generalization. The lowest utility loss by our proposed approach is because of the autonomous ECs of QIs.

Figure 7B shows the query error with respect to varying selectivity for $k = 4$ and dataset = 6,000. With increase in selectivity (*i.e.,* high predicates) less number of records will be selected, which results in a low error rate in the anonymized data. Again, our proposed (θ^*, k)-utility algorithm has continuously low query error as compared to the θ-Sensitive k-Anonymity, (p, l)-angelization, and (k, l)-diversity techniques because of the independent anonymization of the QT with respect to the ST. The low query error in our proposed approach depicts the low difference in the tuples selection between the original and anonymized releases.

Privacy loss

The privacy loss means re-identification of an individual record in the anonymized dataset. In this work, we are using two different methods to measure privacy loss. One is the actual record intersection method, while another is the probability method.

Record intersection

Loss in the privacy of data is the identification of vulnerable records in the anonymized data which ultimately re-identify an individual record. Equation (25) measures an average number of vulnerable records.

The original dataset contains an input file that contains the total data that is not yet anonymized *i.e.,* Table 1, the output table indicates the published tables in the anonymized form, *i.e.,* QT12a and ST12b.

$$VulnerableRecords = Actual \cap Output. \tag{25}$$

The vulnerability of number of records in (k, l)-diversity and(p, l)-angelization is higher than the θ-Sensitive k-Anonymity and (θ^*, k)-utility as shown in Fig. 8A. This is due to the fact that the (k, l)-diversity uses a transaction generalization technique (*Gong et al., 2017*). The lowest common cut on transaction generalization hierarchy leaves the single SA from different sub-set unprotected and vulnerable, as shown in Fig. 2 that causes sVer attack which breaches the privacy of that specific individual. In (p, l)-angelization the vulnerability exists due to the sensitive attributes fingerprint correlation attack, as mentioned in *Khan et al. (2020c)*. As in (p, l)-angelization, the QT and ST tables have one-to-one correspondence (discussed in Motivation Scenario III), so the adversary can correlate both tables and can easily create a single table. In that way, the *ma* and *nma* attacks (see definitions in Section 3) can be performed on the dataset. The variance-based privacy implementation by θ-Sensitive k-Anonymity is stronger. However for the 1:M dataset, it becomes useless to achieve privacy because in such data each record consists of more than

one tuple, and attackers can perform *sva*, and *csa* attacks on the dataset. However, the proposed (θ^*, k)-utility further improves its privacy by categorizing the SA into categorical SA values (*i.e.,* Fig. 2), and prevents all such attacks (*i.e., ma*, *nma*, *sva*, *csa*, and *sVer*). This reduces the re-identification of an individual record and provides more data protection as compared to its counterparts, as shown in Fig. 8A.

Record linkability

In this subsection, the impact of the privacy parameter record linkability (RL) is analyzed through our proposed (θ^*, k)-utility algorithm in comparison to the (k, l)-diversity, (p, l)-angelization and θ-Sensitive k-Anonymity. RL is a measure of disclosure risk (*i.e.,* privacy loss) and is the probability of correctly linked records between the original and the anonymized data. For a record $t_i \in$ T, record linkage probability in anonymized form $P_{RL}(t_i^*)$, is calculated using Eq. (26).

$$P_{RL}(t_i^*) = \begin{cases} \dfrac{1}{|EC_j|} & : t_i \epsilon EC_j \\ 0 & : otherwise \end{cases} \tag{26}$$

where EC_j is generalized group of records in QT with minimum distance from t_i. The RL for complete microdata T is then calculated in Eq. (27) as below.

$$RL = \sum_{t_{i \epsilon T}} P_{RL}(t_i^*). \tag{27}$$

For example, record $t_i \in$ T is put into EC_1 after anonymization. Now for each original record t_i find the closest EC in anonymized QT, let say it is EC_2. If EC_1 is EC_2, the record t_i is linked and is computed *via* Eq. (26). We finally sum RL of all original records as shown in Eq. (27). Figure 8B shows the privacy loss (*i.e.,* RL) for the proposed (θ^*, k)-utility, in comparison to the (k, l)-diversity, (p, l)-angelization, and θ-Sensitive k-Anonymity algorithms. The lower value of RL shows the lower privacy loss and vice versa. The probability to link a record from the microdata T is high with an anonymized EC of small size, because the intruder already knows the QIs of the intended individual, and a record can easily be linked with a few number of records group. So the privacy loss with respect to RL for small k is high as compared to high value of k.

Figure 8B shows that the highest privacy loss in (k, l)-diversity, is due to the lowest common cut on the sensitive values (*i.e., sVer* attack) which becomes vulnerable because of the remaining sub-set of sensitive values unprotected (see Fig. 2). In (p, l)-angelization, the fingerprint correlation attack and the one-to-one correspondence between the QT and ST are the due reasons of privacy losses. In θ-Sensitive k-Anonymity the attackers can easily perform *sva* and *csa* attacks because of its inapplicability for 1:M type of data. Figure 8B depicts that the RL in all these approaches is high for small value of k, because the probability of linking the target quasi identifier values with small number of records is high as compared to the larger size EC. However, our proposed (θ^*, k)-utility algorithm not only categorizes the SA values to implement privacy in sensitive values but also contributes in the form of QS-loose linkability for implementing privacy both in QT and ST, and implementing utility in the QT only. The QS-loose linkability not only minimizes the

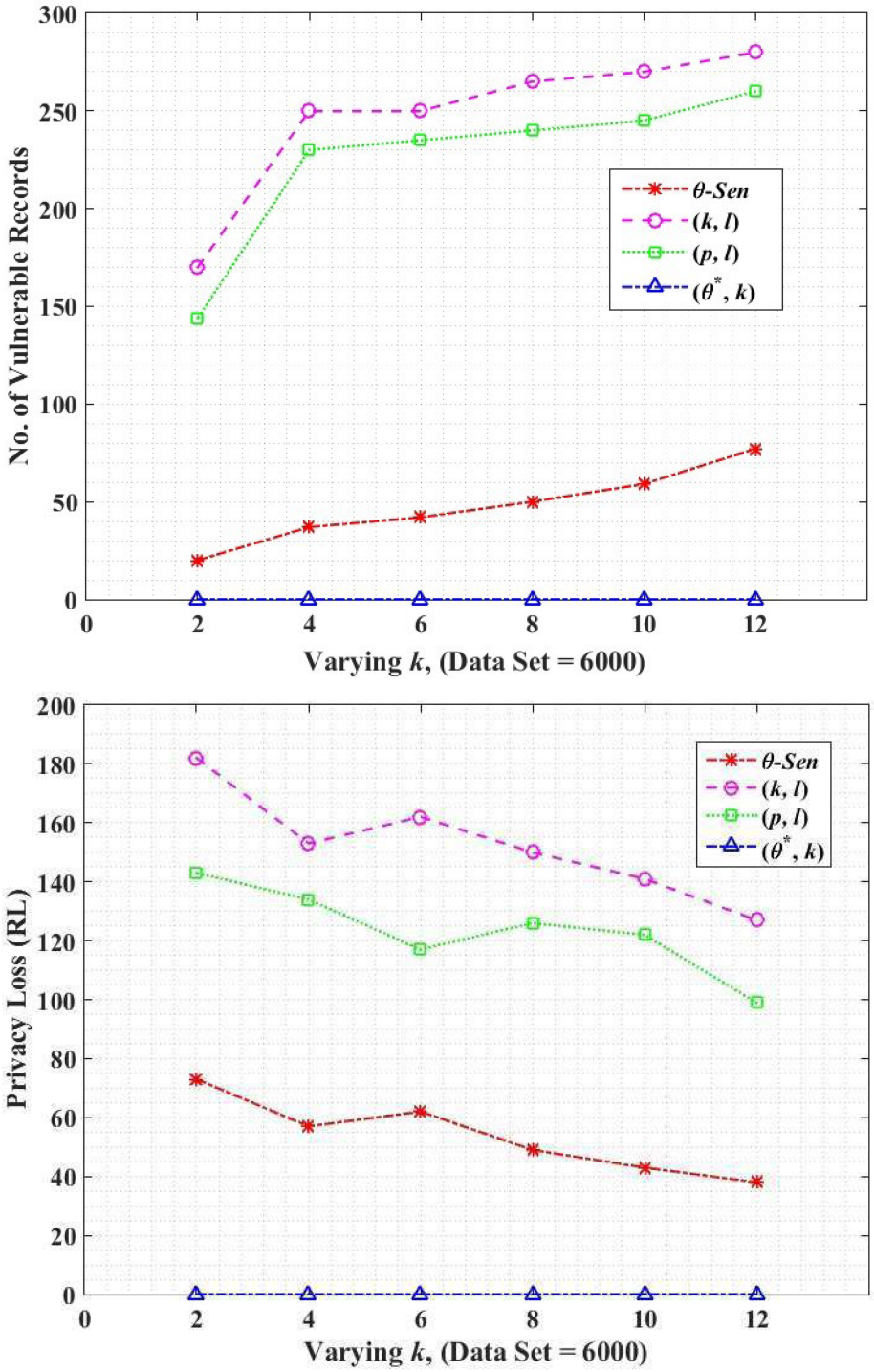

Figure 8 (A) Record intersection; (B) record linkability.

Full-size DOI: 10.7717/peerjcs.1255/fig-8

chances of record linkability but almost vanishes it. So, the beauty in the novelty of the proposed (θ^*, k)-utility algorithm is the contribution both for privacy and utility at the same time.

Execution time

Computational efficiency is the overall execution time of an algorithm. Figure 9 shows the execution time of our proposed (θ^*, k)-utility algorithm along with its counterpart techniques. In both Figures *i.e.,* Figs. 9A and 9B, the (p, l)-angelization has the highest execution time because of the weight calculation and handling 1:M-MSA data. The proposed (θ^*, k)-utility algorithm has higher execution time than θ-Sensitive k-Anonymity, because of the additional work of categorising the SA and creating one-to-many loose linkability between QT and ST, along with the variance calculations for SAs. The (k, l)-diversity has the lowest execution time because of the simple approach of the algorithm *i.e.,* only 1:M generalization and splitting the attributes.

DISCUSSION

The results show that the proposed (θ^*, k)-utility algorithm outperforms all its compared counterparts concerning utility and privacy. Our proposed (θ^*, k)-utility algorithm for measuring NCP, independently generalizes the QI values to create less distance ECS for any size of k, which results in low utility loss compared to its counterparts. For the query accuracy, the lowest utility loss by our proposed approach is because of the autonomous ECs on QIs. The (θ^*, k)-utility improves privacy by categorizing the SA into categorical SA values also using the variance-based privacy implementation. The proposed (θ^*, k)-utility algorithm has higher execution time than θ-Sensitive k-Anonymity and (k, l)-diversity because of the additional work along with the variance calculations and has lower execution time than (p, l)-angelization. So our proposed (θ^*, k)-utility algorithm is best to achieve higher data privacy and data utility as compared to its counterparts.

CONCLUSION AND FUTURE WORK

This article addresses the problem of anonymizing the 1:M microdata with significantly improving the utility of anonymized release. We proposed an anonymization algorithm which prevent any possible attack *e.g.*, membership attack (ma), non-membership attack (nma), sensitive variance attack (sva), categorical similarity attack (csa), and sensitive vertical attack (sVer), which may exists in either θ-Sensitive k-Anonymity (*Khan et al., 2020*), or in (k, l)-diversity (*Gong et al., 2017*), or in (p, l)-angelization (*Kanwal et al., 2019*) techniques. The proposed solution; (θ^*, k)-utility, extends the applicability of θ-Sensitive k-Anonymity (*Khan et al., 2020*), for anonymizing the 1:M microdata. The (θ^*, k)-utility algorithm executes by taking three proactive steps: transformation, sensitive bucket creation, and quasi generalized buckets creation. The SA values *i.e.,* COVID symptoms, in ST are categorized into Table 2 CtgT *i.e.,* Low, Mild, High, Severe, and A-symptomatic, for the purpose to implement privacy in the Table ST Table 12B. The QS-loose linkability between the Table 12A QT and Table 12B ST, not only implements privacy in both tables but also significantly improves the utility of the anonymized data. The results from

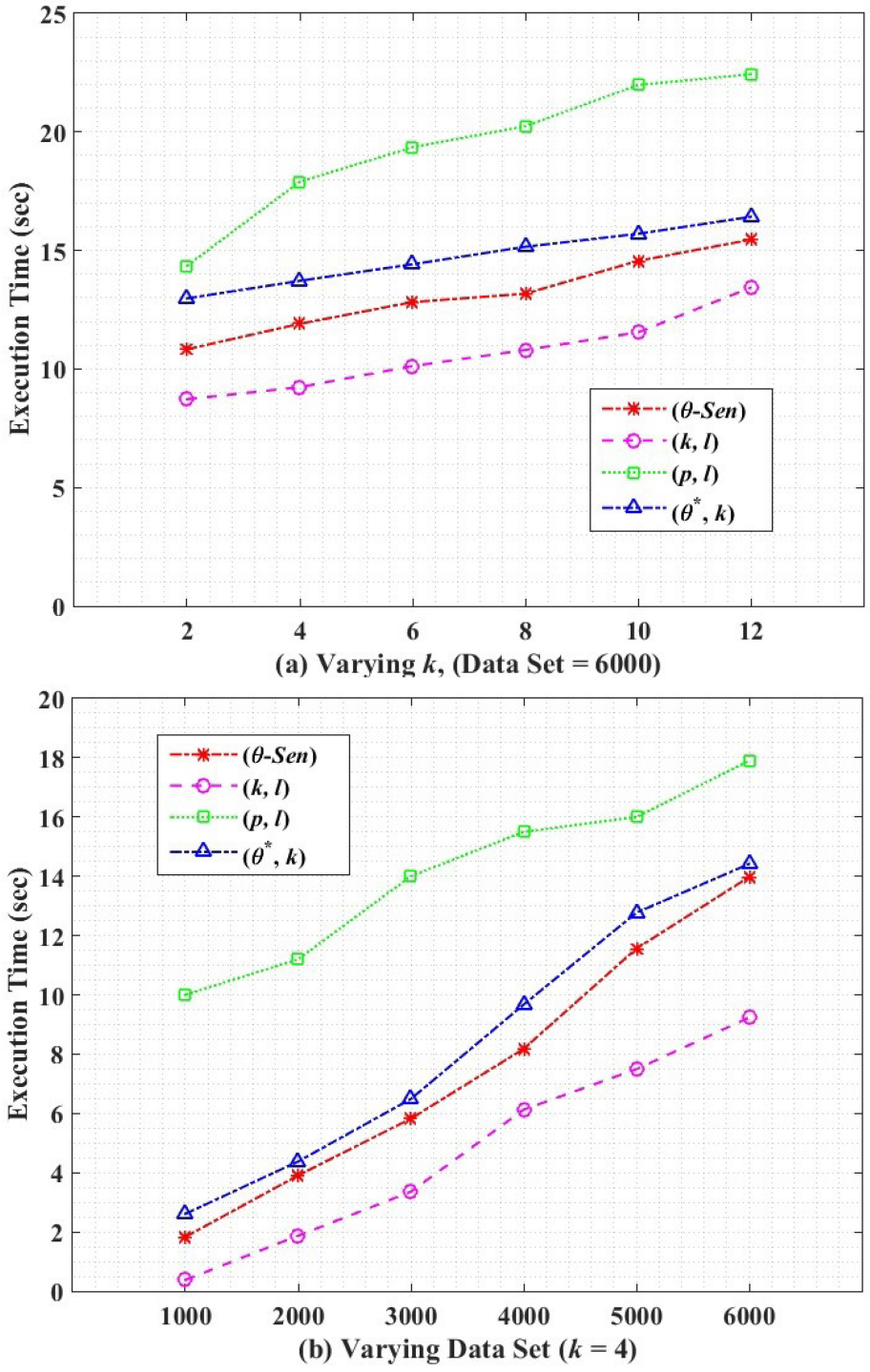

Figure 9 Execution time.

Full-size DOI: 10.7717/peerjcs.1255/fig-9

experiments depicts that with respect to both utility and privacy the proposed (θ^*, k)-utility algorithm outperforms all its compared counterparts.

For future work considerations, the proposed algorithm can be extended to implement privacy in a dynamic data publishing scenario (*Xiao & Tao, 2007*; *Khan et al., 2020b*) for periodic or non-periodic updates. Similarly, the proposed work can be extended to a cluster based anonymization technique to more efficiently overcome the problem of privacy and utility paradigm (*Safi & Hwang, 2022*). Another privacy extension can be privacy-preserving federated learning (PPFL) (*Yin, Zhu & Hu, 2021*). PPFL is a collaborative training process based on iterative model averaging where the user generated data is not directly shared with any third party which greatly benefits the used data from not being disclosed to any un-identified intruder.

ADDITIONAL INFORMATION AND DECLARATIONS

Funding

This work was supported by TU Wien Bibliothek through its Open Access Funding Programme. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Grant Disclosures

The following grant information was disclosed by the authors:
TU Wien Bibliothek through its Open Access Funding Programme.

Competing Interests

The authors declare there are no competing interests.

Author Contributions

- Rabeeha Fazal conceived and designed the experiments, performed the experiments, performed the computation work, prepared figures and/or tables, authored or reviewed drafts of the article, and approved the final draft.
- Razaullah Khan conceived and designed the experiments, performed the experiments, performed the computation work, prepared figures and/or tables, authored or reviewed drafts of the article, and approved the final draft.
- Adeel Anjum conceived and designed the experiments, analyzed the data, performed the computation work, authored or reviewed drafts of the article, and approved the final draft.
- Madiha Haider Syed conceived and designed the experiments, analyzed the data, authored or reviewed drafts of the article, and approved the final draft.
- Abid Khan conceived and designed the experiments, analyzed the data, authored or reviewed drafts of the article, and approved the final draft.
- Semeen Rehman conceived and designed the experiments, analyzed the data, authored or reviewed drafts of the article, and approved the final draft.

Data Availability

The following information was supplied regarding data availability:

The data set used in the work is derived from the publicly available dataset available at the UC Irvine Machine Learning Repository: https://archive.ics.uci.edu/ml/datasets/adult.

The derived data set is available at Zenodo: Ronny Kohavi, & Barry Becker. (1996). UCI Machine Learning- Adult Dataset [Data set]. Zenodo. https://doi.org/10.5281/zenodo.7214275.

Supplemental Information

Supplemental information for this article can be found online at http://dx.doi.org/10.7717/peerj-cs.1255#supplemental-information.

REFERENCES

Al-Khafajiy M, Baker T, Chalmers C, Asim M, Kolivand H, Fahim M, Waraich A. 2019. Remote health monitoring of elderly through wearable sensors. *Multimedia Tools and Applications* **78(17)**:24681–24706 DOI 10.1007/s11042-018-7134-7.

Al-Khafajiy M, Webster L, Baker T, Waraich A. 2018. Towards fog driven IoT healthcare: challenges and framework of fog computing in healthcare. In: *Proceedings of the 2nd international conference on future networks and distributed systems*. 1–7.

Albulayhi K, Tošić PT, Sheldon FT. 2020. G-Model: a novel approach to privacy-preserving 1: M microdata publication. In: *2020 7th IEEE international conference on cyber security and cloud computing (CSCloud)/2020 6th IEEE international conference on edge computing and scalable cloud (EdgeCom)*. Piscataway: IEEE, 88–99.

Amin Z, Anjum A, Khan A, Ahmad A, Jeon G. 2022. Preserving privacy of high-dimensional data by l-diverse constrained slicing. *Electronics* **11(8)**:1257.

Anjum A, Malik SuR, Choo K-KR, Khan A, Haroon A, Khan S, Khan SU, Ahmad N, Raza B. 2018. An efficient privacy mechanism for electronic health records. *Computers & Security* **72**:196–211 DOI 10.1016/j.cose.2017.09.014.

Anjum A, Farooq N, Malik SUR, Khan A, Ahmed M, Gohar M. 2019. An effective privacy preserving mechanism for 1:M microdata with high utility. *Sustainable Cities and Society* **45**:213 DOI 10.1016/j.scs.2018.11.037.

Choudhury O, Gkoulalas-Divanis A, Salonidis T, Sylla I, Park Y, Hsu G, Das A. 2019. Differential privacy-enabled federated learning for sensitive health data. ArXiv preprint. arXiv:1910.02578.

Dang LM, Piran MdJ, Han D, Min K, Moon H. 2019. A survey on internet of things and cloud computing for healthcare. *Electronics* **8(7)**:768 DOI 10.3390/electronics8070768.

Fazal R, Shah MA, Khattak HA, Rauf HT, Al-Turjman F. 2022. Achieving data privacy for decision support systems in times of massive data sharing. *Cluster Computing* 1–13.

Gong Q, Luo J, Yang M, Ni W, Li X-B. 2017. Anonymizing 1: M microdata with high utility. *Knowledge-Based Systems* **115**:15–26 DOI 10.1016/j.knosys.2016.10.012.

Iyer R, Rex R, McPherson KP, Gandhi D, Mahindra A, Singh A, Raskar R. 2021. Spatial K-anonymity: a privacy-preserving method for COVID-19 related geospatial technologies. ArXiv preprint. arXiv:2101.02556.

Jain P, Gyanchandani M, Khare N. 2020. Improved k-anonymize and l-diverse approach for privacy preserving big data publishing using MPSEC dataset. *Computing and Informatics* **39(3)**:537–567 DOI 10.31577/cai_2020_3_537.

Jayapradha P, Alotaibi Y, Khalaf OI, Alghamdi SA. 2022. Heap bucketization anonymity an efficient privacy-preserving data publishing model for multiple sensitive attributes. *IEEE Access* **10**:28773–28791 DOI 10.1109/ACCESS.2022.3158312.

Kanwal T, Anjum A, Malik SUR, Sajjad H, Khan A, Manzoor U, Asheralieva A. 2021. A robust privacy preserving approach for electronic health records using multiple dataset with multiple sensitive attributes. *Computers & Security* **105**:102224 DOI 10.1016/j.cose.2021.102224.

Kanwal T, Shaukat SAA, Anjum A, Choo K-KR, Khan A, Ahmad N, Ahmad M, Khan SU , et al. 2019. Privacy-preserving model and generalization correlation attacks for 1: M data with multiple sensitive attributes. *Information Sciences* **488**:238–256 DOI 10.1016/j.ins.2019.03.004.

Khan R, Tao X, Anjum A, Kanwal T, Malik SUR, Khan A, Rehman WU, Maple C. 2020a. θ-Sensitive k-Anonymity: an anonymization model for IoT based electronic health records. *Electronics* **9(5)**:716 DOI 10.3390/electronics9050716.

Khan R, Tao X, Anjum A, Malik SR, Yu S, Khan A, Rehman W, Malik H. 2020b. (τ, m)-slicedBucket privacy model for sequential anonymization for improving privacy and utility. *Transactions on Emerging Telecommunications Technologies* **33(6)**:e4130.

Khan R, Tao X, Anjum A, Sajjad H, Malik SUR, Khan A, Amiri F. 2020c. Privacy preserving for multiple sensitive attributes against fingerprint correlation attack satisfying c-diversity. *Wireless Communications and Mobile Computing* **2020**:1–18.

Lee YJ, Lee KH. 2017. Re-identification of medical records by optimum quasi-identifiers. In: *2017 19th international conference on advanced communication technology (ICACT)*. Piscataway: IEEE, 428–435.

Liu F, Li T. 2018. A clustering-anonymity privacy-preserving method for wearable iot devices. *Security and Communication Networks* **2018**:1–18.

Lv Z, Piccialli F. 2021. The security of medical data on internet based on differential privacy technology. *ACM Transactions on Internet Technology* **21(3)**:1–18.

Majeed A, Ullah F, Lee S. 2017. Vulnerability-and diversity-aware anonymization of personally identifiable information for improving user privacy and utility of publishing data. *Sensors* **17(5)**:1059 DOI 10.3390/s17051059.

Malik SUR, Khan SU, Srinivasan SK. 2013. Modeling and analysis of state-of-the-art VM-based cloud management platforms. *IEEE Transactions on Cloud Computing* DOI 10.1109/TCC.2013.3.

Michalas A. 2019. The lord of the shares: combining attribute-based encryption and searchable encryption for flexible data sharing. In: *Proceedings of the 34th ACM/SIGAPP symposium on applied computing*. 146–155.

Moonsamy W, Singh S. 2022. Digital vaccination records: exploring stakeholder perceptions in Gauteng, South Africa. *The African Journal of Information and Communication* **29**:1–26.

Müftüoğlu Z, Kızrak MA, Yıldırım TB. 2022. Data sharing and privacy issues arising with COVID-19 data and applications. In: *Data Science for COVID-19*. Cambridge: Academic Press, 61–75.

Nasir M, Anjum A, Manzoor U, Balubaid MA, Ahmed M, Khan A, Ahmad N, Alam M , et al. 2017. Privacy preservation in skewed data using frequency distribution and weightage (FDW). *Journal of Medical Imaging and Health Informatics* **7(6)**:1346–1357 DOI 10.1166/jmihi.2017.2206.

Raju NVSL, Seetaramanath MN, Rao PS. 2019. A novel dynamic KCi-slice publishing prototype for retaining privacy and utility of multiple sensitive attributes. *International Journal of Information Technology and Computer Science* **11(4)**:18–32.

Safi , Hwang S. 2022. Toward privacy preservation using clustering based anonymization: recent advances and future research outlook. *IEEE Access* **10**:1–1 DOI 10.1109/ACCESS.2022.3175219.

Shahzad A, Lee YS, Lee M, Kim Y-G, Xiong N. 2018. Real-time cloud-based health tracking and monitoring system in designed boundary for cardiology patients. *Journal of Sensors* **2018**: *Available at https://www.hindawi.com/journals/js/2018/3202787/*.

Sheikhtaheri A, Tabatabaee Jabali SM, Bitaraf E, TehraniYazdi A, Kabir A. 2022. A near real-time electronic health record-based COVID-19 surveillance system: an experience from a developing country. *Health Information Management Journal* **2022** DOI 10.1177/18333583221104213.

Song F, Ma T, Tian Y, Al-Rodhaan M. 2019. A new method of privacy protection: random k-anonymous. *IEEE Access* **7**:75434–75445 DOI 10.1109/ACCESS.2019.2919165.

Sun W, Cai Z, Li Y, Liu F, Fang S, Wang G. 2018. Security and privacy in the medical internet of things: a review. *Security and Communication Networks* **2018**:1–9.

Tu Z, Zhao K, Xu F, Li Y, Su L, Jin D. 2018. Protecting trajectory from semantic attack considering {k}-Anonymity, {l}-Diversity, and {t}-Closeness. *IEEE Transactions on Network and Service Management* **16(1)**:264–278.

Wang J, Li H, Guo F, Zhang W, Cui Y. 2019a. D2D big data privacy-preserving framework based on (a, k)-anonymity model. *Mathematical Problems in Engineering* **2019**:1–11.

Wang J, Du K, Luo X, Li X. 2019b. Two privacy-preserving approaches for data publishing with identity reservation. *Knowledge and Information Systems* **60(2)**:1039–1080 DOI 10.1007/s10115-018-1237-3.

Xiao X, Tao Y. 2006. Anatomy: simple and effective privacy preservation. In: *Proceedings of the 32nd international conference on very large data bases*. 139–150.

Xiao X, Tao Y. 2007. M-invariance: towards privacy preserving re-publication of dynamic datasets. In: *Proceedings of the 2007 ACM SIGMOD international conference on Management of data*. 689–700.

Yaseen S, Abbas SMA, Anjum A, Saba T, Khan A, Malik SUR, Ahmad N, Shahzad B, Bashir AK. 2018. Improved generalization for secure data publishing. *IEEE Access* **6**:27156–27165 DOI 10.1109/ACCESS.2018.2828398.

Ye Y, Liu Y, Wang C, Lv D, Feng J. 2009. Decomposition: privacy preservation for multiple sensitive attributes. In: *Database Systems for Advanced Applications: 14th International Conference, DASFAA 2009, Brisbane, Australia, April 21–23, 2009. Proceedings*. Berlin Heidelberg: Springer, 486–490.

Yin X, Zhu Y, Hu J. 2021. A comprehensive survey of privacy-preserving federated learning: a taxonomy, review, and future directions. *ACM Computing Surveys (CSUR)* **54(6)**:1–36.

Zhang L, Xuan J, Si R, Wang R. 2017. An improved algorithm of individuation k-anonymity for multiple sensitive attributes. *Wireless Personal Communications* **95(3)**:2003–2020 DOI 10.1007/s11277-016-3922-4.