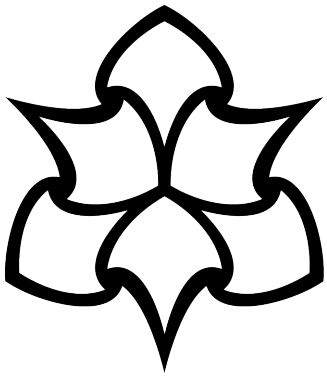

IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

SPECIAL SECTION ON CYBER-THREATS AND COUNTERMEASURES IN THE HEALTHCARE SECTOR

Received February 28, 2018, accepted March 31, 2018, date of publication May 7, 2018, date of current version June 5, 2018.

Digital Object Identifier 10.1109/ACCESS.2018.2828398

Improved Generalization for Secure Data Publishing

SABA YASEEN[1], SYED M. ALI ABBAS[1], ADEEL ANJUM[1], TANZILA SABA[2], ABID KHAN[1], SAIF UR REHMAN MALIK[1], NAVEED AHMAD[1], BASIT SHAHZAD[3], AND ALI KASHIF BASHIR[4]

[1]Department of Computer Sciences, COMSATS University Islamabad, Islamabad 44000, Pakistan
[2]College of Computer and Information Sciences, Prince Sultan University, Riyadh 11586, Saudi Arabia
[3]Department of Computer Science, National University of Modern Languages, Islamabad 44000, Pakistan
[4]Department of Science and Technology, University of the Faroe Islands, Tórshavn 100, Faroe Islands

Corresponding author: Naveed Ahmad (naveedahmad@comsats.edu.pk)

This work was supported by the Machine Learning Research Group, Prince Sultan University Riyadh, Saudi Arabia under Grant RG-CCIS-2017-06-02. The authors are grateful for this support.

ABSTRACT In data publishing, privacy and utility are essential for data owners and users respectively, which cannot coexist well. This incompatibility puts the data privacy researchers under an obligation to find newer and reliable privacy preserving tradeoff-techniques. Data providers like many public and private organizations (e.g. hospitals and banks) publish microdata of individuals for various research purposes. Publishing microdata may compromise the privacy of individuals. To prevent the privacy of individuals, data must be published after removing personal identifiers like name and social security numbers. Removal of the personal identifiers appears as not enough to protect the privacy of individuals. K-anonymity model is used to publish microdata by preserving the individual's privacy through generalization. There exist many state-of-the-arts generalization-based techniques, which deal with pre-defined attacks like background knowledge attack, similarity attack, probability attack and so on. However, existing generalization-based techniques compromise the data utility while ensuring privacy. It is an open question to find an efficient technique that is able to set a trade-off between privacy and utility. In this paper, we discussed existing generalization hierarchies and their limitations in detail. We have also proposed three new generalization techniques including conventional generalization hierarchies, divisors based generalization hierarchies and cardinality-based generalization hierarchies. Extensive experiments on the real-world dataset acknowledge that our technique outperforms among the existing techniques in terms of better utility.

INDEX TERMS Generalization hierarchies, K-anonymity, distortion ratio, global/local recoding.

I. INTRODUCTION

V Data publishing by various officialdom sets the stage for the data users to conduct extensive researches with different determinations. For example, banks publish their data for analysis, so that economists analyze the data and make decisions accordingly. Hospitals publish their data for world health organizations and pharmaceutical researchers. During current era, data publishing is obligatory for analysts and researchers. It is a prerequisite for making decisions and further developments in various fields. The publishable data has sensitive and confidential information about the individuals (i.e. data owners) along with the personally identifiable and quasi-identifiers information. Data publishing in its original form is an open threat [1]–[3] to individual's privacy like generalization and suppression, incognito, On-the-fly hierarchies, Improved on-the-fly hierarchies, and Top-down specialization.

Generalization replaces the original values with specific general values, realistic to the original value. It is useful for data analysis and various research purposes when less-specific data is required. Generalization hierarchies can be used for Data warehousing, Data mining, Machine learning and Object-oriented databases [4].

However, old-growing generalization (recoding) concept/methods are getting importance day by day and the impact is outshining as the data about individuals (microdata) is also growing exponentially into different dimensions. To overcome the current privacy-preservation challenges in diverse dimensions, most of the researchers working in different domains have proposed state-of-the-art frameworks based on generalization methods [18]–[21].

In this paper, we discuss existing generalization hierarchies and their limitations. To overcome these limitations, we propose three novel techniques of generalization hierarchies.

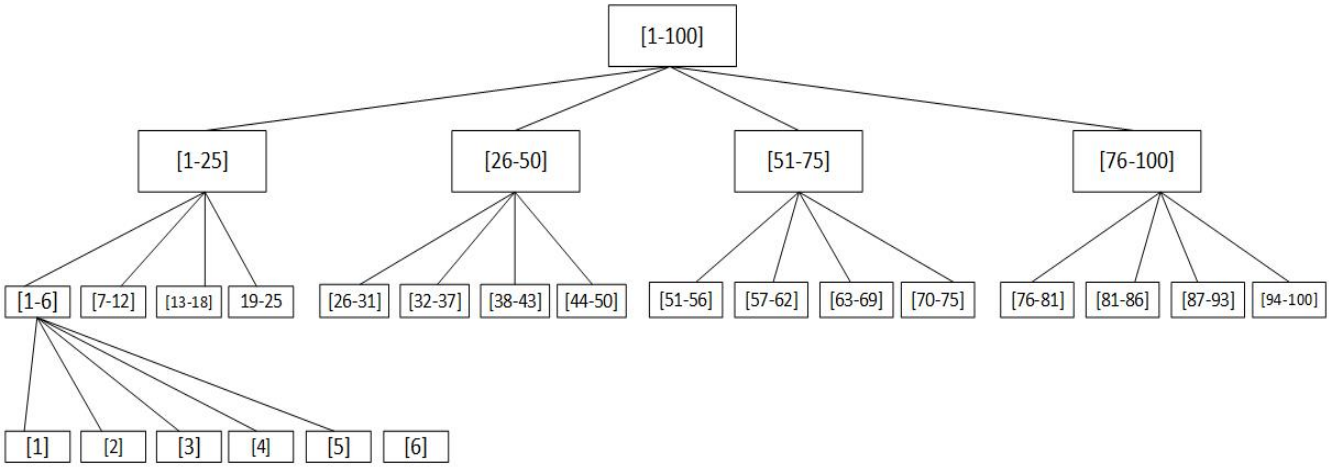

FIGURE 1. Generalization hierarchy.

Experiments on the real-world dataset 'ADULT' (included in machine learning repositories on UCI machine Irvine) advocates that our proposed techniques outperform among all the existing generalization hierarchies.

In order to go deep into the existing generalization hierarchies, understanding of basic definitions and preliminaries is a pre-requisite.

Definition 1 (Attributes): "*Let a table t* $(a1, \ldots an)$ *which contains finite set of attributes, and attributes of 't' are* $(a1, \ldots an)$" we have 14 attributes (Age, Work class, Final Weight, Education, Education-num, Marital status, Occupation, Relationship, Race, Sex, Capital-gain, Capital-Loss, Hours-per-week, and Native-country).

Definition 2 (Quasi-Identifiers): "*Let a table t* $(a1, \ldots an)$. *A quasi-identifier (QI) of t is a set of attributes* $\{a1, \ldots ak\} \subseteq \{a1, \ldots an\}$ *which are commonly used to be shared or published.*" For example, in our ADULT Dataset Age, Education and Hours-per-week are Quasi-identifiers.

Definition 3 (k- Anonymous Dataset): "*Let a table t* $(a1, \ldots an)$ *and* QI_t *be the quasi-identifiers which are a subset of t. t achieves k-anonymity if each sequence of data values in* $t[QI_t]$ *repeats with at least k time.*"

Definition 4 (Recoding/Generalization): "*The concept of replacing exact values of a QI by general values in such a way that actual value resides between general value.*" For example, exact "age = 12" can be replaced with "age groups" like age = [1 to 20].

Definition 5 (Local recoding): "In this type of Recoding, all values of the QI are generalized to the same level of taxonomy." For example [1-20], [21-30], [31-40]...

Definition 6 (Global Recoding): "*In this type of Recoding, some values of a single QI may be more generalized as compared to some of its other values.*" For example [1-20], [21-50], [51-60].

Definition 7 (Generalization Hierarchy): "*A taxonomy that represents different possibilities for generalizing a QI.*" For example, age attribute can be represented as $([1 - 20])$ or $([1 - 10], [11 - 20])$ *or* $([1 - 5], [6 - 10], [11 - 15], [16 - 20])\}$.

Definition 8 (Generalization Lattice): "*A taxonomy that represents all possible combinations of nodes in generalization hierarchies.*"

Definition 9(Distortion Ratio): "*The amount of data utility that has been sacrificed for anonymity.*" Let the height of the generalized attribute 'A_i' is 'H_{ij}' for the tuple 't_j'. Distortion of all attributes in a generalized data set is equal to the sum of all values in a generalized data set. $\sum i, j = h_{i,j}$

Distortion ratio is equal to the distortion of the generalized dataset divided by the distortion of the fully generalized dataset [7].

Table 1 shows generalized data having three attributes age, local and global recoding. In local recoding, all values of quasi-identifiers have been generalized to same level of taxonomy. For example [1-20], [21-40], [41-60]. Some values of quasi-identifier may be more generalized as compared to some of its values in global recoding. For example [1-20], [21-30], [31-60].

TABLE 1. Local vs global recoding.

Age	Global Recoding	Local Recoding
12	1-10	1-10
23	11-20	11-20
25	21-50	21-30
35	51-60	31-40
18	61-90	41-50

Generalization techniques are based on two different types of attributes; categorical and numerical. The categorical attribute is the one, which can take limited or fixed value, and the numerical attribute is the one, which can take any value within a range. $K-$minimal generalization, full domain generalization, incognito, top-down, predefined

TABLE 2. Comparison of existing generalization method.

Technique Name	Problem domain	Attribute Type	Recoding Model	Drawback	Attacks deals with
Generalization and Suppression [1]	• Used to provide privacy to the individuals	Categorical attributes	Global recoding	• Difficult for the publisher • Easy to re-identify user after release	• Unsorted matching attack • Temporal attack • complementary attack
K-minimal generalization and suppression [2]	• Used to provide privacy to the individuals • Ability to link is difficult	Categorical attributes	Global recoding	• This is not applicable for all generalization techniques	• Re-identification attacks • Background knowledge attack
Full Domain Generalization Hierarchy [9]	• Achieve privacy • Compromise utility	Categorical attributes	Global recoding	• Compromise utility • Information loss	NA
Top down specialization [10]	• Maintains privacy VS utility • Generalized table while preserving its usefulness • Some data is useless after generalization	Categorical attributes	Global recoding	• Performance is slow	Background knowledge attacks
Incognito [11]	• Incognito is based on global recoding • Used for minimal full domain generalization • Performance is fast as compared to all previous algorithm	Categorical attributes	Global recoding	• Not scalable • may generate more distortion	Linking attacks
Pre-Defined Generalization Hierarchy [12]	• construct before data masking • usually done by users	Numerical attributes	Local recoding	• less flexible • produce same results for categorical and numerical attributes	
Cell Level generalization [13]	• gives different values for each tuple	Numerical Attribute	Local Recoding	• a lot of information loss	NA
Hierarchy Free Model [8]	• done during anonymization process • minimize information loss	Numerical attributes	Local recoding	• Requires pre-existing hierarchies for generalization	NA
OTF [4]	• Previous Generalization algorithm may lose information • Automatically generates generalization hierarchies	Numerical attributes	Local recoding	• Very less information loss as compared to the previous algorithm	NA
IOTF [14]	• Privacy violation • Severity of privacy violation • Caters all attributes in Generalization hierarchy	Numerical attributes	Local recoding	• It does not construct k-anonymous and l-diverse generalization hierarchies	NA
DCHT [15]	• Dynamically create generalization hierarchies • Perform better than OTF and IOTF	Numerical attributes	Local Recoding	• It does not produce better results where data set is large • No. of nodes at each level depends on no. of tuples each data set have	NA

hierarchies, hierarchy-free model, on-the-fly (OTF) hierarchy and improved-on-the fly (IOTF) hierarchies are existing generalization techniques. Among them, OTF and IOTF only deal with numerical attributes and the rest deal with categorical attributes. We overview the related techniques and their limitations in the next section.

II. BACKGROUND

A. K-MINIMAL GENERALIZATION

Sweeney and Latanya [2] presented k-minimal generalization and suppression. K-minimal generalization has been based on global recoding model; it is difficult for the attacker to apply linking attacks and gain access to the individual's personal information. However, an attacker may use background knowledge for information retrieval. Furthermore, K-minimal generalization is not applicable to all generalization hierarchies. Same generalization techniques have been used in [19].

B. FULL DOMAIN GENERALIZATION HIERARCHY

Full domain generalization deals with categorical attributes. It maps attributes to a more general domain in domain generalization hierarchy. Full domain generalization achieves privacy but compromises the data utility [8], [9], [22].

C. TOP-DOWN SPECIALIZATION

Fung *et al.* [10] put-forwarded Top-Down specialization. It was based on global recoding model for generalization hierarchies. Though it maintains an effective balance between privacy and utility, its efficiency is not up to the mark. There are certain situations, in which it may encounter background knowledge attacks.

D. INCOGNITO

LeFevre *et al.* [11] designed a new model incognito for domain generalization hierarchies. It was based on global recoding model. Incognito's performance was fast among all the previous generalization hierarchies. However, Incognito was not scalable and might leak information by applying linking attack. The updated version has been published in [25].

E. PRE-DEFINED GENERALIZATION HIERARCHY

Pre-defined generalization hierarchies are also known as Iyengar's model. Usually, users construct them before data masking. They are used as they are, we can not alter them for generalization. It is a less flexible technique that produces same generalization hierarchies for both numerical attributes and categorical attributes [12], [23].

F. CELL LEVEL GENERALIZATION

Lyengar's model has been extended to cell level generalization. It allows mapping the values with different generalized values. Its only drawback is giving different generalized values for each tuple that leads to huge information loss [13], [22].

G. HIERARCHY FREE MODEL

Hierarchies are constructed according to the decision being taken during anonymization [12]. Hierarchy free model is a more supple generalization technique that helps in minimizing the information loss. Its only limitation is that k-anonymity model requires pre-existing hierarchies for numerical attributes [8], [24].

H. ON-THE-FLY DOMAIN GENERALIZATION HIERARCHY

Campan *et al.* [4] and Kim *et al.* [25] proposed on-the-fly domain generalization hierarchy. It was based on local recoding model. It prevents information loss during generalization and suppression process. It may encounter background knowledge attack.

I. IMPROVED ON-THE-FLY GENERALIZATION HIERARCHY

Campan *et al.* [14] improved the performance of on-the-fly generalization hierarchy. The improved technique prevents data privacy violations and information loss. It does not construct k-anonymous and l-diverse generalization hierarchies.

J. DCHT

In [15], DCHT were introduced. DCHT considers cell level generalization for creating hierarchies. They are very complex to create as we consider values only in Table 1. We need to check all the values from the Table 1 prior to set a range. In large datasets, it appears to be a very complex task.

We present a comparison of existing generalization hierarchies in this paper. The limitations and attacks dealt by these techniques have also been listed in this table. K-minimal generalization [2] deals with linkage attack and background knowledge attack while generalization and suppression [1] deal with the unsorted matching attack, temporal and complementary attacks. Top-down specialization [10] deals with background attack. Incognito tries to fix the privacy breach by linkage attack.

III. PROPOSED METHOD

We performed extensive experiments on the real-world dataset 'ADULT's dataset' published by UCI. The dataset contains 48834 records in total. The details of the attributes, included in our dataset have been depicted in Table 3. We selected set of three attributes as quasi-identifiers that are *Age*, *Education*, and *Work Hours*. The domain of these attributes can be seen in Table 3. Afterwards, we used three different methods for creating generalization hierarchies for each of these attributes. Then, we produced nine generalization hierarchies (three quasi-identifiers and three methods). Furthermore, we produced three generalization lattices (one for each of the three methods, we used for creating generalization hierarchies). Later, we computed results (value of "k" and generalization cost) for each node of our three generalization lattices. Finally, we compared those results by plotting the value of "k" against the generalization cost. The cost of generalization was measured in terms of the distortion produced by each node.

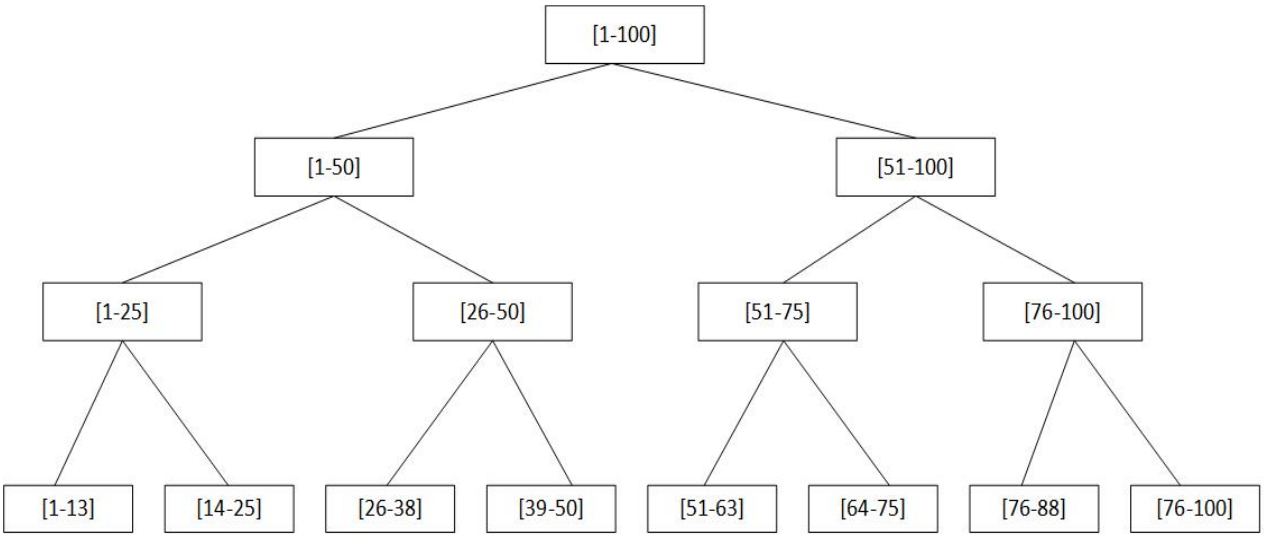

FIGURE 2. Conversational based generalization.

TABLE 3. Adults dataset information.

Attribute Name	Domain:
Age	'Min Value =17; **Max** Value = 90
Work class	Private, Self-emp-not-inc, Self-emp-inc, Federal-gov etc
Final Weight	Min Value =12285; **Max** Value = 1490400
Education	Bachelors, Some-college, 11th, HS-grad, Prof-school,Assoc-acdm, Assoc-voc, 9th, 7th-8th, 12th, Masters, 1st-4th, 10th, Doctorate, 5th-6th, Preschool.
Education-num	Mm Value =1; Max Value = 16
Marital-status	Married-civ-spouse, Divorced, Never-married, Separated etc
Occupation	Tech-support, Craft-repair, Sales, Exec-managerial etc
Relationship	Wife, Own-child, Husband, Not-in-family, Unmarried etc
Race	White, Asian-Pac-Islander, Amer-Indian-Eskimo, Other, Black.
Sex	Female, Male.
Capital-gain	Min Value 0; Max Value = 99999
Capital-loss	Mm Value ˙.:1; Max Value = 4356
Hours-per-week	Mm Value =1; Max Value = 99
Native-country	United-States, Cambodia, England, Puerto-Rico, Canada etc

We proposed three methods including *Conventional Generalization Hierarchies* (CGH), *Divisors Based Generalization Hierarchies* (DBGH) and *Cardinality-Based Generalization Hierarchies* (CBGH) for creating the generalization hierarchy. Our proposed solution guarantees preservation of more data utility on comparing with existing state of the art IOTF.

A. METHOD 1 (CONVENTIONAL GENERALIZATION HIERARCHIES - CGH)

In our first method, we started from the highest generalized level and created lower level generalizations by splitting each parent interval into two equivalent child intervals. For example, in case of work hours, we started from the largest interval [1 to 100] and created its child nodes by splitting the whole interval into two equivalent intervals, i.e. [1 to 50] and [51 to 100]. Then, in the next lower level, each of these two intervals was further split into two equals; thus, the four intervals [1 to 25], [26 to 50], [51 to 75], and [76 to 100] were formed. This process continued until we found the level with smallest desired intervals. The hierarchies, we obtained using this method, have been depicted in Figure 2. In this case, the distortion ratio (generalization cost) has been calculated as under:

i. The highest distortion (associated with the root of the generalization hierarchy/tree), is equal to the height of the tree.
ii. Distortion of a node "x" (where x is any node except the root) is equal to "distortion of the parent of x" / 2.

B. METHOD 2 (DIVISORS BASED GENERALIZATION HIERARCHIES – DBGH)

In our second method, we created as many levels we could create intervals of the whole range. Like, in the case of work hours the whole range was [1 to 100]. Therefore, we found the following possible intervals:

i. [1 to50 and 51 to 100]
ii. [1 to 25, 26 to 50, 51 to 75, and 76 to 100]

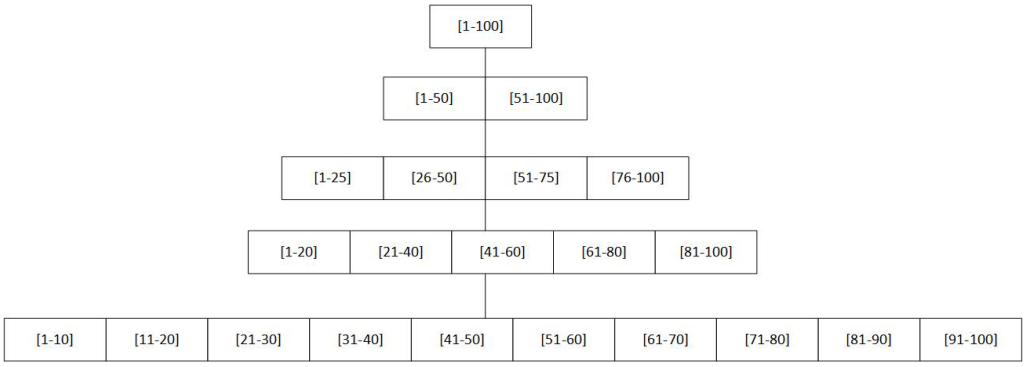

FIGURE 3. Divisor based generalization.

Algorithm 1 For Divisor Based Generalization Hierarchy Creation

Create Divisor Based Hierarchy (Max Value, Min Value)
Inputs: Max Value /* Maximum value in the QI */, Min Value /* Minimum value in the QI */
Output: Generalization hierarchy (T) for the QI
1. If MaxValue is Prime then MaxValue = MaxValue + 1
2. T = NULL
3. Find all possible divisors of the MaxValue and store in the array *DivList
4. For (i=1; i<=DivList.size; i++)
Loop
Lc = CreateLevel (DivList[i]) /*Create a new level (Lt) of generalization tree with interval size DivList[i]*/
If T == NULL then T = Lt Else Add Lt at top of T
End Loop; S. Return T;
CreateLevel (DivList[i])
1. Lc = NULL
2. For (j=1; j<MaxValue; j=j+DivList[i])
Loop
If (j+DivList[i] < MinValue) Then Continue; /*Skip iteration*/
Add the node at level (Lt) : j —> j+DivList[i]
End Loop;
3. Return Lt;

iii. [1 to 20, 21 to 40, 41 to 60, 61 to 80, and 81 to 100]
iv. [1 to 10, 11 to 20, ... , 81 to 90, and 91 to 100]

Although, lower level intervals were possible, our desired minimum interval was 10, so we ignored those levels. The hierarchies formed by this method can be seen in Figure 3. In this case, distortion of the root node was, once again, equal to the height of the tree. However, the calculation of lower levels-distortions was based on interval length of each level, using the proportions formula.

Algorithm 2 For Cardinality-Based Generalization Hierarchy Creation

Create Cardinality-Based Hierarchy (MaxValue, Min-Value)
Inputs: MaxValue /* Maximum value in the QI V, Min-Value /* Minimum value in the QI */
Output: Generalization hierarchy (T) for the QI
1. If MinValue is ODD and MaxValue is EVEN or MinValue is EVEN and MaxValue is ODD then
MaxValue MaxValue + 1
2. TL = CreateLeafLevel (MaxValue, MinValue)
3. T=TL;
4. Repeat TN = CreateLevel (TL)
Add TN at top of T; TL = TN;
Until TN . No_of Nodes = 1;
5. Return T;
CreateLeafLevel (MaxValue, MinValue)
1. TL = NULL
2. For (i=MinValue; i<MaxValue; i=i+2)
Loop
Add the node at level (TL) : 1+2
End Loop;
3. Return TL;
CreateLevel (TL)
1. If TL . No_of Nodes is EVEN then
Create the level TN by merging nodes of TL as: N1 + N2, N3 + N4 N (n − 1) + N(n)
Else
Assign Cardinalities (no of tuples) to each node at level TL
Create the level TN by merging (n−1) nodes of TL with each other, but leave one node untouched.
Do it according to the following rules:
i. Only consecutive nodes can be merged.
ii. Choose the option that produces nodes at TN with minimized cardinality differences.
2. Return TN;

C. METHOD 3 (CARDINALITY-BASED GENERALIZATION HIERARCHIES – CBGH)

Our third method has been based on the concept of local recoding or multi-valued generalization. In this method, same

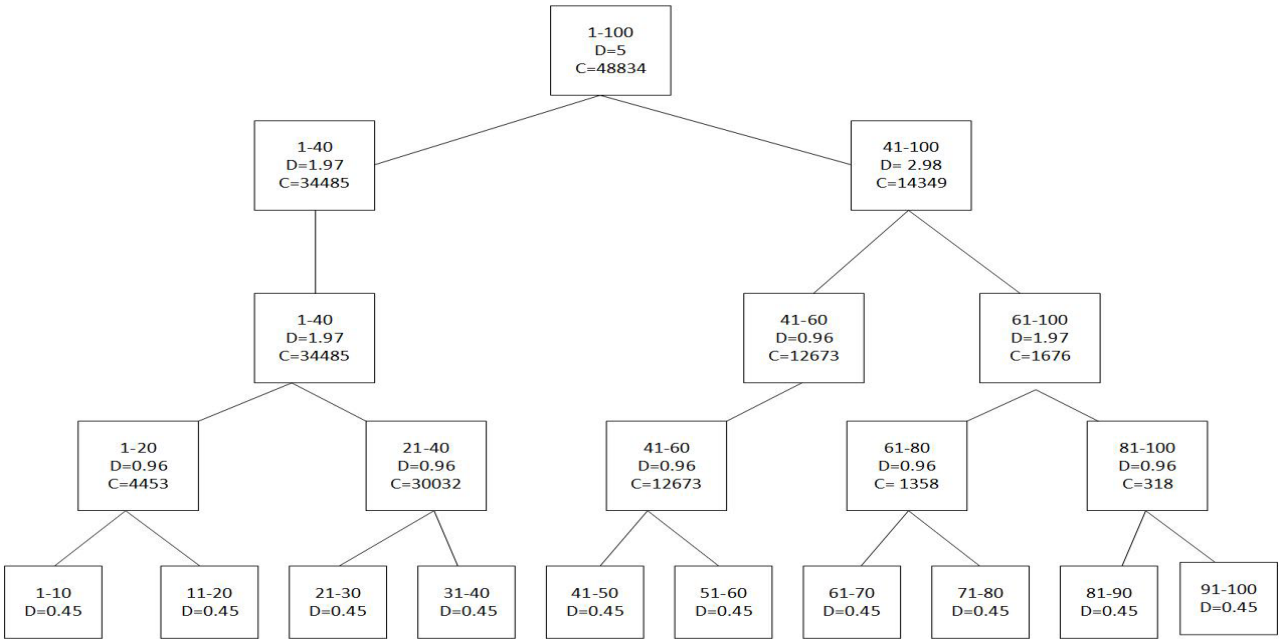

FIGURE 4. Cardinality-based generalization.

level nodes-distortion does not necessarily to be the same. However, different nodes at the same level may have different distortions. The reason behind is that the same level nodes-intervals may be different from each other. In this method, we constructed tree from bottom to top by merging lower level intervals to form the larger intervals. Each of the parent nodes was constructed by merging its lower level nodes. If there were four nodes at any level i, its parent level (means $i + 1$) would be consisting of two nodes (formed by merging node1 with node2 and node3 with node4). However, if a level had three nodes, then parent level would again be having two nodes; one could be constructed by merging two of the lower nodes where the second would be same as the child level node. Now the question is that what two nodes would be merged. Several tuples (called the cardinality) in each node would decide that what to be kept same at the parent level. The nodes having low cardinalities would be merged, so the same level nodes-cardinality difference might be minimized.

IV. RESULTS AND DISCUSSION

Generalization is one of the leading methods for data anonymization. Many current anonymization frameworks have been based on different generalization methods [24], [26]–[28]. However, it needs to have more practice to know that how to apply any of the generalization technique for the specific dataset. Each attribute in the dataset has its own domain and range. It is necessary to set the balanced ranges for generalizing the records. The basic motivation of our work is to anonymize the dataset in such a way that the high data utility still prevails. The least change in medical records can lead to producing wrong results [29]. Global recoding has been used extensively for publishing dynamic data in [30]. It highly necessary to ensure high data utility after data anonymization.

We did an imperative study on generalization based privacy solutions [1], [2] and its different variants [4], [14], [15]. The three new techniques CGH, DBGH, CBGH have been introduced to improve the performance (data utility) of existing generalization techniques.

Comparison Table 2 has been populated to evaluate the performance of existing techniques of generalization. In Table 4, we presented a comparison of our proposed techniques with IOTF and proved that our proposed techniques outperformed.

The performance of these hierarchies has been based on three observations. We compared the proposed method with the existing state-of-the-art **IOTF** based on these observations. Detail discussion and results were discussed below:

A. OBSERVATION BASED ON NO. OF NODES AT EACH LEVEL

It has been observed that the level 2 and 3 of generalization hierarchies IOTF have a maximum number of nodes. Number of nodes means that we can encounter more information loss. When we observed the hierarchies according to no. of nodes at each level, the conventional method appeared to be the best one because it produced very less number of nodes and thus ensured privacy. Cardinality-based generalization hierarchies appeared to be second best hierarchies based on the no. of nodes. Figure 5 depicts the comparison of IOTF with our proposed techniques.

B. OBSERVATION BASED ON DISTORTION RATIO

It has been observed that Cardinality and divisor based generalization hierarchies had highest distortion ratio at level 5.

TABLE 4. IOTF generalization & distortion ratio comparing with proposed methods.

Generalization Technique	IOTF	Conventional Method	Divisor Based Generalization Hierarchy	Cardinality-based Generalization Hierarchy
Maximum height of Generalization Tree	4	4	5	5
Distortion Ratio				
Level 1	0.041	0.5	0.1	0.45
Level 2	0.25	1	0.125	0.96
Level 3	1	2	0.625	0.96, 1.97
Level 4	4	4	2.5	1.97, 2.98
Level 5	NA	NA	5	5
Max. No. of Nodes at each Level				
Level 1	24	8	10	10
Level 2	16	4	5	5
Level 3	4	2	4	3
Level 4	1	1	2	2
Level 5	NA	NA	1	1

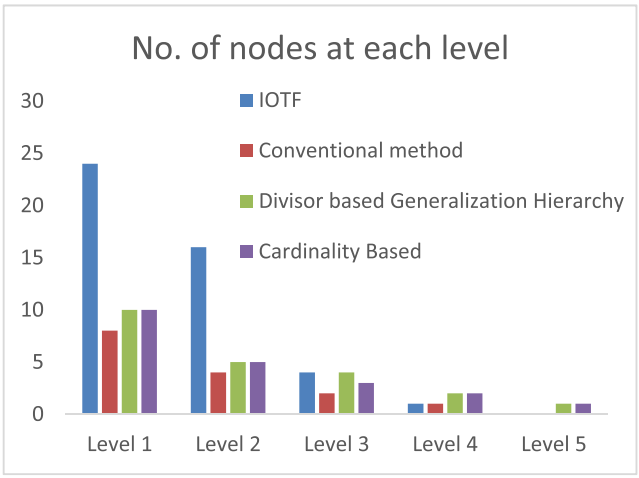

FIGURE 5. No. of Nodes At Each Level Comparison of IOTF with proposed solutions.

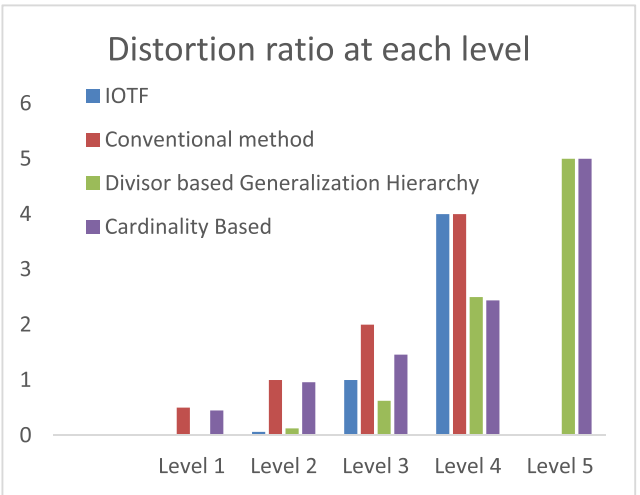

FIGURE 6. Distortion Ratio Comparison Of IOTF With Proposed Solutions.

At level 4, 3, 2, 1-distortion ratio decreased twice or thrice as compared to IOTF and Conventional method. It was clearly observed that cardinality based generalization hierarchies had minimum distortion ratio. Figure 6 shows the results of distortion ratio at each level for all the techniques. Table 4 provides an insight of exact values of distortion ratio at each level.

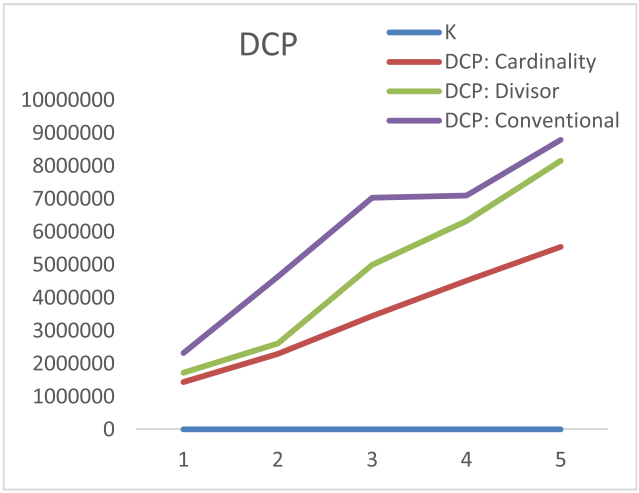

FIGURE 7. DCP Comparison Of Proposed Solutions With No. of K.

C. OBSERVATION BASED ON DCP

Discernibility Penalty [16], [17] is a method of measuring the quality of generalization hierarchies. The minimum value of DCP shows the effectiveness of generalization hierarchy. Cardinality-based generalization hierarchy produces minimum DCP as compared to Divisor based and the conventional method. Figure 7 depicts the above-mentioned result. Here '*k*' is the number of groups shown vertically.

Figure 5 shows the number of nodes at each level for IOTF, Conventional and Cardinality-based generalization hierarchies. Figure 6 and Table 4 show distortion ratios at each level. In above results, we compared our proposed solution with existing state-of-the-art and showed that our solution outperformed substantially better among existing solutions in terms of data utility.

V. CONCLUSION

In this paper, existing generalization techniques, their limitations and the methods to overcome the limitations have been discussed. Three new techniques including CGH, DBGH, and CBGH have been proposed. The performance of these hierarchies has been based on three observations; a number of nodes at each level, distortion ratio and DCP (Discernibility Penalty). Results & discussion section transparently compared the newly proposed techniques with the existing state-of-the-art IOTF. Conventional (CGH) method was observed ensuring privacy by limiting the number of nodes at each level as compared to the other generalization hierarchies. Cardinality-based generalization hierarchy (CBGH) appeared as having minimum distortion ratio, as well as Discernibility Penalty. Putting all these facts under consideration, it can be said that CBGH is best amongst all the newly proposed and existing techniques.

As a future direction, following improvements can be made:

i. Intelligent mechanism to produce generalization hierarchies and set optimum ranges according to the dataset automatically.
ii. Generating generalization hierarchies and sharing big data in the distributed framework.
iii. Generating generalization hierarchies and set ranges in Textual datasets.

REFERENCES

[1] L. Sweeney, "k-anonymity: A model for protecting privacy," *Int. J. Uncertainty, Fuzziness Knowl.-Based Syst.*, vol. 10, no. 5, pp. 557–570, 2002.
[2] L. Sweeney, "Achieving k-anonymity privacy protection using generalization and suppression," *Int. J. Uncertainty, Fuzziness Knowl.-Based Syst.*, vol. 10, no. 5, pp. 571–588, 2002.
[3] L. Sweeney, "Achieving k-anonymity privacy protection using generalization and suppression," *Int. J. Uncertainty, Fuzziness Knowl.-Based Syst.*, vol. 10, no. 5, pp. 571–588, 2002.
[4] A. Campan and N. Cooper, "On-the-fly hierarchies for numerical attributes in data anonymization," in *Proc. Workshop Secure Data Manage.*, 2010, pp. 13–25.
[5] J. Han and Y. Fu, "Dynamic generation and refinement of concept hierarchies for knowledge discovery in databases," in *Proc. KDD Workshop*, 1994, pp. 157–168.
[6] W. W. Chu and K. Chiang, "Abstraction of high level concepts from numerical values in databases,"in *Proc. KDD Workshop*, 1994, pp. 133–144.
[7] K. Patel, S. Parameswaran, and S. L. Shee, "Ensuring secure program execution in multiprocessor embedded systems: A case study," in *Proc. 5th Int. Conf. Hardw./Softw. Codesign Syst. Synth.*, 2007, pp. 57–62.
[8] K. LeFevre, D. J. DeWitt, and R. Ramakrishnan, "Mondrian multidimensional k-anonymity," in *Proc. 22nd Int. Conf. Data Eng. (ICDE)*, Apr. 2006, p. 25.
[9] P. Samarati, "Protecting respondent's identities in microdata release," *IEEE Trans. Knowl. Data Eng.*, vol. 13, no. 6, pp. 1010–1027, Nov./Dec. 2001.
[10] B. C. M. Fung, K. Wang, and P. S. Yu, "Top-down specialization for information and privacy preservation," in *Proc. 22nd Int. Conf. Data Eng. (ICDE)*, Apr. 2005, pp. 205–216.
[11] K. LeFevre, D. J. DeWitt, and R. Ramakrishnan, "Incognito: Efficient full-domain k-anonymity," in *Proc. ACM SIGMOD Int. Conf. Manag. Data*, 2005, pp. 49–60.
[12] V. S. Iyengar, "Transforming data to satisfy privacy constraints," in *Proc. 8th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2002, pp. 279–288.
[13] M. Lunacek, D. Whitley, and I. Ray, "A crossover operator for the *k*-anonymity problem," in *Proc. 8th Annu. Conf. Genetic Evol. Comput.*, 2006, pp. 1713–1720.
[14] A. Campan, N. Cooper, and T. M. Truta, "On-the-fly generalization hierarchies for numerical attributes revisited," in *Proc. Workshop Secure Data Manage.*, 2011, pp. 18–32.
[15] V. K. Vatsavayi and S. K. Adusumalli, "Cost effective dynamic concept hierarchy generation for preserving privacy," *J. Inf. Knowl. Manage.*, vol. 13, no. 4, p. 1450035, 2014.
[16] T. Iwuchukwu and J. F. Naughton, "K-anonymization as spatial indexing: Toward scalable and incremental anonymization," in *Proc. 33rd Int. Conf. Very Large Data Bases*, 2007, pp. 746–757.
[17] K. V. Ramana and V. V. Kumari, "Graph-based local recoding for data anonymization," *Int. J. Database Manage. Syst.*, vol. 5, no. 4, p. 1, 2013.
[18] A. H. M. SarowarSattar, J. Li, X. Ding, J. Liu, and M. Vincent, "A general framework for privacy preserving data publishing," *Knowl.-Based Syst.*, vol. 54, pp. 276–287, Dec. 2013.
[19] H. Zakerzadeh, C. C. Aggarwal, and K. Barker, "Privacy-preserving big data publishing," in *Proc. 27th Int. Conf. Sci. Stat. Database Manage.*, 2015, Art. no. 26.
[20] A. Anjum and G. Raschia, "BangA: An efficient and flexible generalization-based algorithm for privacy preserving data publication," *Computers*, vol. 6, no. 1, p. 1, 2017.
[21] Z. Yan, P. Zhang, and A. V. Vasilakos, "A survey on trust management for Internet of Things," *J. Netw. Comput. Appl.*, vol. 42, pp. 120–134, Jun. 2014.
[22] 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, p. 1103, 2014.

[23] X. Xiao, G. Wang, and J. Gehrke, "Differential privacy via wavelet transforms," *IEEE Trans. Knowl. Data Eng.*, vol. 23, no. 8, pp. 1200–1214, Aug. 2011
[24] Q. Gong, J. Luo, M. Yang, W. Ni, and X.-B. Li "Anonymizing 1:M microdata with high utility," *Knowl.-Based Syst.*, vol. 115, pp. 15–26, Jan. 2017.
[25] S. Kim, H. Lee, and Y. D. Chung, "Privacy-preserving data cube for electronic medical records: An experimental evaluation," *Int. J. Med. Inform.*, vol. 97, pp. 33–42, Jan. 2017.
[26] M. Terrovitis, G. Poulis, N. Mamoulis, and S. Skiadopoulos "Local suppression and splitting techniques for privacy preserving publication of trajectories," *IEEE Trans. Knowl. Data Eng.*, vol. 29, no. 7, pp. 1466–1479, Jul. 2017.
[27] P. Jagwani and S. Kaushik, "Privacy in location based services: Protection strategies, attack models and open challenges," in *Proc. Int. Conf. Inf. Sci. Appl.*, 2017, pp. 12–21.
[28] S. K. Adusumalli and V. V. Kumari, "An efficient and dynamic concept hierarchy generation for data anonymization," in *Proc. Int. Conf. Distrib. Comput. Internet Technol.*, 2013, pp. 488–499.
[29] M. Kamran and M. Farooq "An information-preserving watermarking scheme for right protection of EMR systems," *IEEE Trans. Knowl. Data Eng.*, vol. 24, no. 11, pp. 1950–1962, Nov. 2012.
[30] A. Anjum *et al.*, "T-safety: A privacy model for sequential publication with arbitrary updates," *Comput. Secur.*, vol. 66, pp. 20–39, May 2017.

SABA YASEEN received the master's degree in information security from COMSATS University Islamabad in 2014. Her research interests include data privacy using generalization.

SYED M. ALI ABBAS is currently pursuing the M.S. degree in information security with COMSATS University Islamabad in 2014. His research interests include data privacy using generalization.

ADEEL ANJUM received the Ph.D. degree (Hons.) in 2013. He is currently an Assistant Professor with the Department of Computer Sciences, COMSATS University Islamabad, Islamabad, Pakistan. He has several publications in international conferences. He is also the author of a book on data privacy. His research interests include data privacy using artificial intelligence techniques. He serves in the technical program committees of various international conferences.

TANZILA SABA received the Ph.D. degree from the Faculty of Computing, Universiti Teknologi Malaysia, Malaysia, in 2012, with a focus on document information management and security. She is currently an Eminent Researcher with the Image Processing Research Group. She is also an Assistant Professor with the College of Computer and Information Sciences, Prince Sultan University, Riyadh, Saudi Arabia. She has authored over 30 papers published in high-impact-factor journals. She was selected for a Marquis Who's Who 2012 Award for her excellent research achievements around the globe.

ABID KHAN received the Ph.D. degree from the Harbin Institute of Technology. He is currently an Assistant Professor of computer science with COMSATS University Islamabad, Islamabad. His research interests include security and privacy of cloud computing (outsourced storage and computation), security protocols, digital watermarking, secure provenance, and information systems.

SAIF UR REHMAN MALIK received the Ph.D. degree from North Dakota State University, USA. He is currently an Assistant Professor with COMSATS University Islamabad, Islamabad, Pakistan. His research interests include formal verification analysis and modeling, cyber physical systems, and large scale computing systems.

NAVEED AHMAD received the Ph.D. degree in engineering design from the University of Cambridge, England. He joined COMSATS University Islamabad (CIIT) in 2011. He is currently an Assistant Professor with the Computer Science Department, CIIT. He is a member of Research Groups on software engineering and mobile application development, CIIT. His research interests include change management, process management, risk management, and enterprise information systems in particular enterprise resource planning software.

BASIT SHAHZAD has 14 years' research and teaching experience and was with King Saud University, Saudi Arabia, and COMSATS University Islamabad, Islamabad. He is currently with the National University of Modern Languages, Islamabad. He has numerous publications in journals and conferences of international repute and has a very active research profile and a reasonably impressive social impact. His research interests include social network analysis, health informatics, information systems, and software engineering. He values research a lot and his career objectives revolve around conducting and supervising research in information systems (enterprise architecture, software cost and risk modeling, and mitigating technological risks in modern banking), advancements in research methodologies (quantitative, qualitative, mixed method), mobile healthcare, and social networks (theory, empirical social network analysis, micro-blogging-based analysis).

ALI KASHIF BASHIR is currently with the Department of Science and Technology, University of the Faroe Islands. His research interests include computer communications (networks), computer security and reliability, and distributed computing.