K-anonymity Model for Multiple Sensitive Attributes

Nidhi Maheshwarkar
MIT,Ujjain

Kshitij Pathak
MIT, Ujjain

Narendra S. Choudhari
IIT, Indore

ABSTRACT
In today's era acquiring information about others is not difficult task but securing this data form interlopers is a big deal. K-anonymity model used to protect released data. Released data which is available for public used may contain sensitive and non-sensitive data. But K-anonymity model faces changes when set of sensitive attributes are present in the data set. To achieve K-anonymous table with diversity may causes distortion of data in some extent. This paper proposed a new concept to minimize this data distortion without using tuple suppression for M-SA K-anonymity Model.

Keywords
K-anonymity, Attacks on anonymous table, *l*-diversity, multiple sensitive attributes.

1. INTRODUCTION
K-anonymity is the emerging concept for the protection of released data. Anonymity word comes from Greek word 'Anonymia' which means nameless state, without a name or a data or information converted in such form in which no one can infer or predict it. Anonymity typically refers to the state on individual's personal identity or personally identifiable information, being publically unknown.

To convert a data set into a anonymous table traditional approaches are used, before releasing data, data publisher may encrypt or remove some data which causes data disclosure, these attributes are name, surname, Social Security Number etc. Even removing these identifiers data is not secured, and causes linking attack's-anonymity model introduced to control linking attack. When released information linked with confidential table may causes data disclosures. Confidential table contains individual's private data and these tables are generally of any organizations such as hospital or bank etc. Medical status, bank details, property details of an individual may causes severe problem.

K-anonymity model suggest to convert those identifiers (Quasi identifiers, who are responsible for linking attack) in such a manner that adversary does not infer the sensitive information related to them. But it is difficult for a data publisher to generate a anonymous table, when multiple sensitive attributes are present in data set. Sensitive attributes are those attributes which may remain hidden from external usage. These attributes related to individual's medical status, bank details, property details etc.

In the next section we will discuss k-anonymous model, attacks on k-anonymous table and *l*-diversity concept which helps a lot to prevent these attacks. In section 3, we will discuss multiple sensitive attributes and drawbacks of k-anonymous *l*-diverse table in the presence of multiple sensitive attributes. In section 4 we propose an algorithm for M-SA k-anonymous model which helps to protect data from trespasser [1].

2. K-ANONYMITY MODEL AND *l*-DIVERSITY
Many organizations are publishing microdata tables that contain unaggregated information about individuals. These tables contains sensitive and non-sensitive attributes. If the individuals can be uniquely identified in the microdata, then their private information would be disclosed and this is unacceptable. So we need a model to secure sensitive data from intruder.

2.1 K-anonymity
K-anonymity is the emerging concept for database protection. In this approach we converge quasi-identifiers of private table in such a way that adversary can't infer sensitive information related to them and sensitive data remains safe. To convert a normal private table into a secure anonymous table, many techniques such as sampling, swapping values, and adding noise to the data while maintaining some overall statistical properties of the resulting table. However, many uses require release and explicit management of microdata while needing *truthful* information within each tuple. This 'data quality' requirement makes inappropriate those techniques that disturb data and therefore although preserving statistical properties compromise the correctness of single tuple. K-anonymity together with its enforcement via Generalization and Suppression has been therefore proposed as an approach to protect respondent's identities while releasing truthful information [1, 12].

In generalization a value of quasi-identifier is replaced by a less specific and more general value that is faithful to the original. Generalization applies on cell level whereas suppression is performed on tuples. Suppression is hiding of tuples when needed or we can say suppression is not releasing any tuple when causes mismatch to k factor for anonymity. We can generalize a date of birth of an individual in the form of month and year or only year. So this contains some original values as well as increase confusion to adversary to infer sensitive data. Suppression is not performed always. Generally the data publishers ignore to perform suppression because it causes data loss [1].

Let's consider if a private table is to be converted into a protected table, then the data publisher after study decides the minimum value of k such that when the adversary, having the quasi-identifier values, searches for a particular data, he will find minimum k records that only increase more confusion and

S. NO	NONSENSITIVE			SENSITIVE
	ZIP CODE	AGE	NATIONALITY	MEDICAL STATUS
1	456001	28	Russian	Heart Disease
2	130669	29	American	Heart Disease
3	130669	21	Japanese	HIV
4	456001	23	American	HIV
5	148533	50	Indian	Cancer
6	148533	55	Russian	Heart Disease
7	148500	47	American	HIV
8	148500	49	American	HIV
9	456001	31	American	Cancer
10	456001	37	Indian	Cancer
11	130669	36	Japanese	Cancer
12	130669	35	American	Cancer

Fig. 1 Impatient Microdata

NAME	ZIPCODE	AGE	MARITAL STATUS	NATIONALITY
.....	.....	.....	.....	.....
.....	.....	.....	.....	.....
.....	.....	.....	.....	.....
.....	.....	.....	.....	.....
Tom	148533	50	Single	Indian
.....	.....	.....	.....	.....
	.....	.....	.....	.....
.....	.....	.....	.....	.....
.....	.....	.....	.....	.....

Fig 2: Public Table

He can't find the individual's details. But if there is such a record due to which the minimum k factor doesn't matches, we delete or suppress it. This is the concept of Suppression .Figure 1 shows an impatient private table that contains sensitive and non-sensitive attributes. Work as a quasi-identifier that links with external available data and disclosed sensitive information. Figure 2 shows a public table and shows that Tom has Cancer.

Generalization and Suppression are most common and reliable techniques for achieving K-anonymity. Attributes generalization can be performed on dataset may be in two forms, considering each record as a individual or set of records, generally known as domain. Figure 3 show domain and value generalization hierarchies which are applied on domain and individual values both [1, 2].

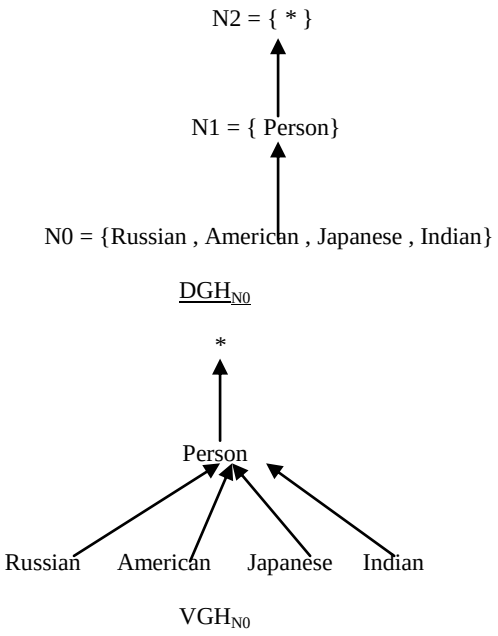

Figure 3. Domain and Value Generalization Hierarchies for NATIONALITY

2.2 Attacks on K-anonymous Table

In section 1, we have already discuss a linking attack and drawbacks related to them, in section 2.1 we discuss K-anonymity model to prevent this model but, attacks are still continued over anonymous table. There are two main major attacks known as Homogeneity and Background Knowledge attacks. Figure 4 shows 4-anonymous impatient microdata. It shows that for any set of quasi-identifiers, there are k-1 same records present in the data set. So interloper doesn't infer individual's sensitive data. But sometimes k-anonymity leads to leakage of information and when this information is disclosed homogeneity attack occurs. Figure 3 shows that for records 9, 10, 11, 12, all persons have cancer, even when this data is anonymized. So there must be some diversity factor that protects anonymized data from adversaries. In diversity, we manage records in a well represented form, which leads to remove leakage of information. In section 2.3, we will discuss *l*-diversity concept in brief. The second type of attack which is hard to remove is Background Knowledge attack. It occurs when an adversary have deep knowledge about individual. Here data publisher have to face problems because a data publisher is unable to know adversaries knowledge about individual. Another problem for data publisher if multiple adversaries are trying to infer a person's information, all have different level of knowledge [4].

2.3 *l*-diversity

Using this background knowledge attack, an adversary can disclose information in two ways: Positive disclosure and Negative disclosure. Wherein positive disclosure, an adversary can correctly identify the value of a sensitive attributes with high probability, while in the negative disclosure the adversary can correctly eliminate some possible values of sensitive attribute with high probability. So after brief study of these attacks we can say that background

knowledge attack is difficult to prevent as compared to homogeneity attack .Diversity ensures that all tuples that share the same values of their quasi-identifiers should have diverse values for their sensitive attributes. Figure shows 3-diverse impatient table. This shows that even if an adversary has background knowledge, there are *l* well represented sensitive values in table. So adversary needs *l*-1 damaging pieces of background knowledge to eliminate *l*-1 possible sensitive values and infer a positive disclosure. Thus by setting the parameter *l*, the data publisher can determine how much protection is provided against background knowledge even if this background knowledge is unknown to the publisher [4].

3. MULTIPLE SENSITIVE ATTRIBUTES

K-anonymity model introduced to protect sensitive attributes from interlopers where sensitive attribute is an attribute whose value for some particular individual must be kept secret from people who have no direct access to the original data. Data publisher needs to prevent privacy disclosure which means someone can simply attack link the publish table T and at least know the individuals suffer from some kinds of privacy disease. This phenomenon is a kind of privacy disclosure [5]. Information disclosure are of three types:

Identity disclosure: An individual is linked to a particular record in the published data.

Attribute disclosure: Sensitive attribute information of an individual is disclosed.

Membership Disclosure: Information about whether an individual's record is in the published data or not is disclosed.

So, the data publisher have to convert a private table in such a manner that if an adversary want to search an individual's identity and have knowledge about quasi-identifiers, finds k-1 records that satisfies k-1 quasi-identifiers. Data publishers have to face problem when multiple sensitive attributes are present in records. Figure 5 shows a table having multiple sensitive attributes [6]. In this table Medical Status, Annual income and occupation are considered as a sensitive attributes. So when a data publisher concentrates to protect one sensitive attributes may cause disclosure of identity due to another one. So we need a technique to control all sensitive attribute. In section 4 we propose an algorithm, which is the extension of [6] prevent multiple sensitive attributes without suppression.

	NONSENSITIVE			**SENSITIVE**
S. NO	**ZIP CODE**	**AGE**	**NATIONALITY**	**MEDICAL STATUS**
1	130***	<30	*	Heart Disease
2	130***	<30	*	Heart Disease
3	130***	<30	*	HIV
4	130***	<30	*	HIV
5	1485**	≥40	*	Cancer
6	1485**	≥40	*	Heart Disease
7	1485**	≥40	*	HIV
8	1485**	≥40	*	HIV
9	130***	3*	*	Cancer
10	130***	3*	*	Cancer
11	130***	3*	*	Cancer
12	130***	3*	*	Cancer

Fig. 4: 4-Anonymous Inpatient Microdata

S.NO.	**ATTRIBUTE**	**TYPE**
1	ZIPCODE	NON-SENSITIVE
2	AGE	NON-SENSITIVE
3	NATIONALITY	NON-SENSITIVE
4	MEDICAL_STATUS	SENSITIVE
5	OCCUPATION	SENSITIVE
6	ANNUAL_INCOME	SENSITIVE

Fig. 5 Description of dataset

4. M-SA K-ANONYMITY MODEL

Algorithm M-SA K-ANONYMITY MODEL
//Temporary is a table of structure like Master (Private Table).Master table has n records and attributes zip code,age,nationality as a non-sensitive attributes and medical status,occupation,annual income as a sensitive attributes.
//Data master table Row 1 to n. F_Trim (a, b) removes b characters from a string Ex F_Trim (123456, 2) = 1234.
//Count_Min_K_Factor_Groupwise (Input_table, Output_ table)will form groups on basis of Nationality, ZIP and Age and return groups data (Nationality, ZIP ,Age) and count of the records per group in the form of table so output data will be (Nationality, ZIP ,Age, Count of the group).
//General_Medical_Status, General_Occupation, General_Annual_Income table contains data of sensitive attributes but having no sensitive information.
//Group_Details table to store group data and count, K anonymity factor (assuming 5), N = Total no of records in Master table.
//Assuming for minimum ZIP length z, on converting 1 to z-1 characters to *; required Anonymity and Diversity will be achieved.
STEP 1: Set Temporary to NULL//Temporary is a table of structure like Master (Private Table).
STEP 2: For Master (i) = 1 to n
{
Do Temporary (i) = Master (i);
}//Copying data from Master Table (having n records) to Temporary
STEP 3: For Temporary (i) = 1 to n
{
Do Temporary (i).Nationality = '*';
}// Updating Nationality to *
STEP 4: For Temporary (i) = 1 to n
{
If Temporary (i).Age <30
{
Do Temporary (i).Age = '<30';
}
Else if Temporary (i).Age =<30 and <= 60
{
Do Temporary (i).Age =' [30-60]';
}
Else
{
Do Temporary (i).Age = '>60 ;}
}
t =0; // t is a variable for adding * to ZIP
STEP 5: t= t+1;
For Temporary (i) = 1 to n Check if marked as a modified go to next record//Only those records are not considered which as shows their status in group _details as a modified.
else

{
Do Temporary (i).Zip = F_Trim (Temporary (i).Zip, t);//Removing last t digits from Zip
For c = 1 to t
{
Temporary (i).Zip = Append (Temporary (i).Zip,'*'); //Adding * at the end of Zip
}
}

STEP6: Count_Min_K_Factor_Groupwise(Temporary,Groups_Details)//Count_Min_K_Factor_Groupwise (Input_table, Output_ table) will form groups on basis of Nationality, ZIP and Age and return groups data (Nationality, ZIP ,Age) and count of the records per group in the form of table so output data will be (Nationality, ZIP ,Age , Count of the group).
p= 0; // p is a variable for General_Medical_Status table
q=0; / / q is a variable for General_Occupation table
r= 0; / / r is a variable for General_ Annual_Income table
j= 1; / / j is a variable for Group_Details table
While Group_Details (j) is not null
{
Do Cnt = Group_Details (j).count;

While Cnt < K

{
Do Temporary (N+1).Nationality= Group_Details (j). Nationality;
Do Temporary (N+1).Zip = Group_Details (j).Zip;
Do Temporary (N+1).Age = Group_Details (j).Age; // Inserting additional similar entries to achieve K Anonymity
p = p+1;
Do Temporary (N+1).Medical_Status= General_Medical_Status (p) // Updating sensitive medical status to non sensitive ex Cancer to Malaria
If p= X //X is number of records in General_Medical_Status table
{
Do p= 0;
}
q= q+1;
Do Temporary (N+1).Occupation= General_Occupation (q) // Updating sensitive Occupation to non sensitive Ex Manager to Clark
If q = Y //Y is number of records in General_Occupation table
{
Do q= 0;
}
r= r+1;
Do Temporary (N+1).Annual_Income = General_Annual_Income (r)// updating sensitive Annual_Income to non sensitive Ex: 2 lakh to 12 lakh
If r = Z //z is number of records in General_Annual_Income table
{ Do r= 0;
} mark all as a modified in Group_Details.
Cnt = Cnt+1;
}
// K Anonymity achieved.
While Cnt > =K
{
Count_Sensetive (Temporary, Group_Details (j), Medical_Status, OP_Medical_Status_Group);
// Count_Sensetive forms groups on basis of sensitive data passed as input parameters (here Medical_Status) and returns a table having two fields Sensitive_Field and its count (Medical_Status and count of Medical_Status (Ex: Cancer , 5))

```
If Group_Details (j). Count = OP_Medical_Status_Group (1). Count
{
        Go to (Step 5);
}

Count_Sensetive (Temporary, Group_Details (j), Occupation, OP_Occupation_Group);
  If Group_Details (j). Count = OP_Occupation_Group (1). Count
{
        Go to (Step 5);
} Count_Sensetive (Temporary, Group_Details (j), Annual_Income, OP_Annual_Income_Group);
If Group_Details (j). Count = OP_Annual_Income_Group (1). Count
{
        Go to (Step 5);
}
 }
         j= j+1;
}
STEP 7: END
```

Figure 6: Algorithm for M-SA K-anonymity model

	NONSENSITIVE			SENSITIVE		
S. NO	ZIP	AGE	NATIONALITY	MEDI_STATUS	OCCUPATION	INCOME
1	13051	26	Indian	Heart Disease	Pilot	3
2	130652	45	American	Heart Disease	Lecturer	4
3	130654	36	Indian	HIV	Distillation Chemists	5
4	13058	29	Indian	Flu	Explosives Handler	5
5	13050	29	Russian	Gastric Ulcer	Doctor	3.5
6	13051	28	Indian	HIV	Doctor	4.5
7	13055	29	American	Flu	QC Inspector (Aircraft)	3.5
8	130660	37	Indian	Cancer	Lecturer	5
9	456001	66	Russian	HIV	Doctor	5
10	456008	61	American	Cancer	Distillation Chemists	3.5
11	130659	31	Indian	Flu	Student	0
12	130665	58	Russian	Cancer	Distillation Chemists	4
13	130654	42	American	Cancer	Lecturer	4
14	130658	50	Russian	Gastric Ulcer	Distillation Chemists	4
15	130661	53	Indian	Cancer	Explosives Handler	5
16	130662	42	Russian	Cancer	Doctor	5
17	130666	31	American	Cancer	Student	0

Impatient Microdata

	NONSENSITIVE			SENSITIVE		
S. NO	Z_CODE	AGE	NATIONALITY	MEDI_STATUS	OCCUPATION	INCOME
1	1305*	<30	*	Heart Disease	Pilot	3
4	1305*	<30	*	Flu	Explosives Handler	5
5	1305*	<30	*	Gastric Ulcer	Doctor	3.5
6	1305*	<30	*	HIV	Doctor	4.5
7	1305*	<30	*	Flu	QC Inspector (Aircraft)	3.5
2	1306**	[30-6	*	Heart Disease	Lecturer	4
3	1306**	[30-6	*	HIV	Distillation Chemists	5
13	1306**	[30-6	*	Cancer	Lecturer	4
14	1306**	[30-6	*	Gastric Ulcer	Distillation Chemists	4
11	1306**	[30-6	*	Flu	Student	0
17	1306**	[30-6	*	Cancer	Student	0
12	1306**	[30-6	*	Cancer	Distillation Chemists	4
15	1306**	[30-6	*	Cancer	Explosives Handler	5
16	1306**	[30-6	*	Cancer	Doctor	5
8	1306**	[30-6	*	Cancer	Lecturer	5
9	45600*	>60	*	HIV	Doctor	5
10	45600*	>60	*	Cancer	Distillation Chemists	3.5
18	45600*	>60	*	Heart Disease	S/w Engg.	2.2
19	45600*	>60	*	Flu	Lecturer	3
20	45600*	>60	*	Gastric Ulcer	S/w Engg.	4

3-SA 5-Anonymous Impatient table

	Satisfies K=5, Height=1, Total Records=5. Where Height refers to the level of generalization.
	Satisfies K=5 (when 3 Records added), Total Records=5.
	Satisfies K=5, Height=2, Total Record=10.
	Sensitive attributes having same values that cause increment in Height.
	Extra added records to satisfy K=5.

Figure 7 Experimental Results

Figure 6 shows Algorithm for M-SA K-anonymity model to protect multiple sensitive attribute. In this model K-anonymity and *l*-diversity achieved. Figure 7 shows the results based on this algorithm. This algorithm also proposed a alternative solution for tuple suppression.Quasi-identifiers are generalized in such a manner it will maintain minimity principle which state that "anonymization method should not generalized, suppress, or distort the data more than it is necessary to achieve k-anonymity." Figure 5 shows a data set description which contains sensitive and non-sensitive attributes. In this algorithm we generalize quasi-identifier age and nationality in fixed level of generalization and we check by increasing level of generalization for zip code to achieve K-anonymity and diversity both.

5. CONCLUSION

This paper proposed an alternative concept for suppression. Suppression means not releasing the attributes which fails to achieve anonymity and *l*-diversity but every record contains individual's details which are unique but when suppression is applied record is ignored which causes data lose which is not a good technique. In M-SA K-anonymity model we add a new record to maintain anonymity as well as diversity. The addition of new records depends upon minimum k factor of the dataset. This is a secure algorithm to maintain usability of data set as well as diversity of the records.

6. REFERENCES

[1] Nidhi Maheshwarkar, Kshitij Pathak, Vivekanand Chourey, "Performance Issues of Various K-anonymity Strategies", International Journal of Computer Technology and Electronics Engineering (IJCTEE), 2011, ISSN No. 2249-6343.

[2] Nidhi Maheshwarkar, Kshitij Pathak, Vivekanand Chourey, "Privacy Issues for K-anonymity Model", International Journal of Engineering Research and Application, 2011, ISSN No. 2248-9622.

[3] V.Ciriani , S. De Capitani di Vimercati , S. Foresti ,P. Samarati,"K-Anonymity",Springer US, Advances In Information Security (2007).

[4] Latanya. Sweeney,"Achieving K-Anonymity privacy protection using generalization and suppression" International journal of Uncertainty,Fuzziness and Knowledge-based Systems,10(5), May 2002,571588.

[5] Pierangela Samarati ,Latanya Sweeney,"Protecting Privacy when Disclosing Information: K-Anonymity and its enforcement through Generalization and Suppression.1998.

[6] A. Machanavajjhala, J.Gehrke,D. Kifer, and M. Venkitasubramaniam. *l*-diversity: Privacy beyond k-anonymity. In Proc.22nd International Conf. Data Engg. (ICDE), page 24 , 2006.

[7] Xinping Hu Zhihui Sun Yingjie Wu Wenyu Hu , Jiancheng Dong " K-Anonymity Based on Sensitive Tuples", 2009 First International Workshop on Database Technology and Applications, 978-0-7695-3604-0/09 /2009 IEEE DOI 10.1109/DBTA.2009.74 M. Young, The Technical Writer's Handbook. Mill Valley, CA: University Science, 1989.

[8] Yingjie Wu, Xiaowen Ruan,Shangbin Liao, Xiaodong Wang," P-Cover K-anonymity model for Protecting Multiple Sensitive Attributes", IEEE,The 5th International Conference onComputer Science & Education Hefei, China. August 24–27, 2010. 978-1-4244-6005-2/10/2010 IEEE.

[9] Rinku Dewri, Indrajit Ray, Indrakshi Ray ,Darrell Whitley," On the Optimal Selection of k in the k–Anonymity Problem".

[10] G. Aggarwal, T. Feder, K. Kenthapadi, R. Motwani, R. Panigrahy, D. Thomas, and A. Zhu. k- Anonymity: Algorithms and hardness. Technical report, Stanford University, 2004.

[11] R. J. Bayardo and R. Agrawal. Data privacy through optimal k - anonymization. In ICDE-2005, 2005

[12] K. LeFevre, D. DeWitt, and R. Ramakrishnan. Incognito: Efficient fulldomain k-anonymity. In SIGMOD, 2005.

[13] A. Meyerson and R. Williams. On the complexity of Optimal k anonymity. In PODS, 2004.

[14] P. Samarati. Protecting respondents' identities in microdata release.In IEEE Transaction s on Knowledge and Data Engineering, 2001.

[15] L. Sweeney. K-anonymity: a model for protecting privacy. International Journal on Uncertainty, Fuzziness and Knowledge-based Systems, 10(5):557–570, 2002.

[16] S. Zhong, Z. Yang, and R. N. Wright. Privacy-enhancing kanonymization of customer data. In PODS, 2005

[17] A. Dobra. Statistical Tools for Disclosure Limitation in Multiway Contingency Tables. PhD thesis , Carnegie Mellon University, 2002

[18] S. L. Kullback and R. A. Leibler. On information and sufficiency. Ann. Math. Stat., 22:79–86, 1951.

[19] Samarati P, Sweeney L (1998). Generalizing data to provide anonymitywhen disclosing information (Abstract). In Proc. of the 17th ACM-SIGMOD-