PASS ALGORITHM

IMPLEMENTATION OF BIG DATA ANONYMIZATION ALGORITHM USING HADOOP FRAMEWORK

Anonymization is one of the main techniques which is being used in recent times to prevent privacy breaches on the published data .one such anonymization technique is k-anonymization technique. The anonymization is a parametric anonymization technique used for data anonymization. The aim of the k-anonymization is to generalize the tuples in way that it can't be identified using quasi identifiers.

In past few years we saw a tremendous growth in data which ultimately led to the concept of the big data .The growth in data made anonymization using conventional processing methods inefficient. To make the anonymization more efficient we used hadoop map reduce to reduce the processing time of anonymization .In this we have divided the whole program into map and reduce parts .More over the data types used in hadoop provide better serialization and transport of data. We performed our experiments on the large data set .The results proved the efficiency of our implementation.

1 INTRODUCTION

1.1 What is Big Data?

Big Data" is data whose scale, diversity, and complexity require new architecture, techniques, algorithms, and analytics to manage it and extract value and hidden knowledge from it…data,

1.2 BIG DATA SIGNIFICANCE IN INDUSTRY AND CHALLENGES

While understanding the value of big data continues to remain a challenge, other practical challenges including funding and return on investment and skills continue to remain at the forefront for a number of different industries that are adopting big data. With that said, a Gartner Survey for 2015 shows that more than 75% of companies are investing or are planning to invest in big data in the next two years. These findings represent a significant increase from a similar survey done in 2012 which indicated that 58% of companies invested or were planning to invest in big data within the next 2 years.

Generally, most organizations have several goals for adopting big data projects. While the primary goal for most organizations is to enhance customer experience, other goals include cost reduction, better targeted marketing and making existing processes more efficient. In recent times, data breaches have also made enhanced security an important goal that big data.

VALUE OF BIG DATA IN ANALYTICS:

big data is more real time in nature than traditional Dataware applications.
Traditional Dataware applications are not suited for big data.
Using big data we can provide more accurate suggestions and solutions .

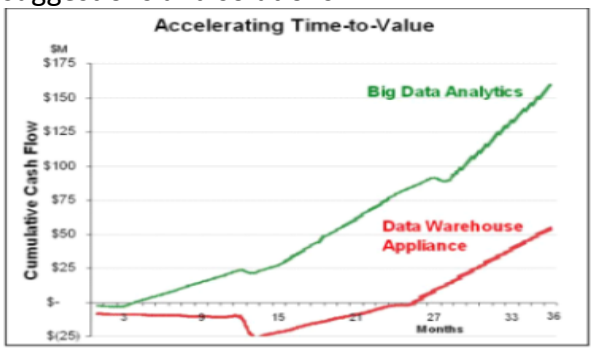
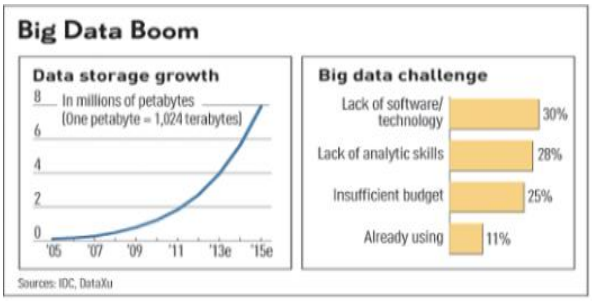


Fig 1

1.3 CHALLENGES IN HANDLING BIG DATA:



THE BOTTLE NECK IS TECHNOLOGY:
new architecture,algorithms techniques are needed.

ALSO IN TECHNICAL SKILLS:
Experts in using the new technology and using big data.

1.4 DATA STREAM

DEFINITION: Big data associated with time stamp is called big data stream.

Examples of data streams:
1 .sensor data
2. call centre records
3. click streams
4. health care data
5. constraints associated with data streams:

Privacy protection: i.e the data streams are extracted from various sources which consist of many individuals hence the sensitive data of any individuals must not be leaked.

Real time processing:
since the data is not static in nature real time processing is required and at present,not many algorithms are there to process the dynamic data

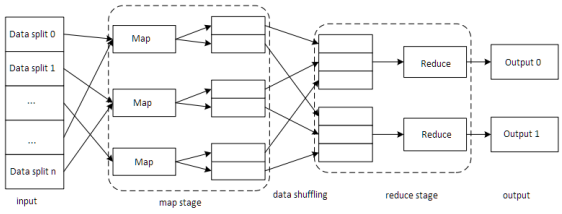
1.5 What is MapReduce?

MapReduce is a processing technique and a program model for distributed computing based on java. The MapReduce algorithm contains two important tasks, namely Map and Reduce. Map takes a set of data and converts it into another set of data, where individual elements are broken down into tuples (key/value pairs). Secondly, reduce task, which takes the output from a map as an input and combines those data tuples into a smaller set of tuples. As the sequence of the name MapReduce implies, the reduce task is always performed after the map job.
The major advantage of MapReduce is that it is easy to scale data processing over multiple computing nodes. Under the MapReduce model, the data processing

primitives are called mappers and reducers. Decomposing a data processing application into mappers and reducers is sometimes nontrivial. But, once we write an application in the MapReduce form, scaling the application to run over hundreds, thousands, or even tens of thousands of machines in a cluster is merely a configuration change. This simple scalability is what has attracted many programmers to use the MapReduce model.

The Algorithm

- Generally MapReduce paradigm is based on sending the computer to where the data resides!
- MapReduce program executes in three stages, namely map stage, shuffle stage, and reduce stage.
    - **Map Stage** : The map or mapper's job is to process the input data. Generally the input data is in the form of file or directory and is stored in the Hadoop file system (HDFS). The input file is passed to the mapper function line by line. The mapper processes the data and creates several small chunks of data.
    - **Reduce Stage** : This stage is the combination of the **Shuffle** stage and the **Reduce** stage. The Reducer's job is to process the data that comes from the mapper. After processing, it produces a new set of output, which will be stored in the HDFS.
- During a MapReduce job, Hadoop sends the Map and Reduce tasks to the appropriate servers in the cluster.
- The framework manages all the details of data-passing such as issuing tasks, verifying task completion, and copying data around the cluster between the nodes.
- Most of the computing takes place on nodes with data on local disks that reduces the network traffic.
- After completion of the given tasks, the cluster collects and reduces the data to form an appropriate result, and sends it back to the Hadoop server.



2.0 ANONYMIZATION

Anonymization:

Generally the main theme Data anonymization is the use of one or more techniques designed to make it impossible or at least more difficult to identify a particular individual from stored data related to them.

purpose of data anoymization:

1.prevent the privacy of individuals who shared data for various surveys.

2.to implement effective techniques to prevent security breach.

techniques implemented are :

1. encryption
2. hashing
3. generalization
4. supresssion of data
5 .destroy data quality
6.adding mathematical noise

2.1 Few privacy prevention techniques used for static data:

2.1.1 .Computer Security:

Access control and authentication ensure that right people has right authority to the right object at right time and right place.

That's not what we want here. A general doctrine of data privacy is to release all the information as much as the identities of the subjects (people) are protected.

2. 1.2 K-Anonymity:

What is K-Anonymity?

If the information for each person contained in the release cannot be distinguished from at least k-1 individuals whose information also appears in the release.

Ex.

If you try to identify a man from a release, but the only information you have is his birth date and gender. There are k people meet the requirement. This is k-Anonymity.

classification of attributes:

Key Attribute: Name, Address, Cell Phone

which can uniquely identify an individual directly Always removed before release.

Quasi-Identifier: 5-digit ZIP code,Birth date, gender

A set of attributes that can be potentially linked with external information to re-identify entities

87% of the population in U.S. can be uniquely identified based on these attributes, according to the Census summary data in 1991.

How attacks take place:

Hospital data:

Voter data :

DOB	Sex	Zipcode	Disease
1/21/76	Male	53715	Heart Disease
4/13/86	Female	53715	Hepatitis
2/28/76	Male	53703	Brochitis
1/21/76	Male	53703	Broken Arm
4/13/86	Female	53706	Flu
2/28/76	Female	53706	Hang Nail

Name	DOB	Sex	Zipcode
Andre	1/21/76	Male	53715
Beth	1/10/81	Female	55410

carol	10/1/44	Female	90210
Dan	2/21/84	Male	02174
Ellen	4/19/72	Female	02237

from above tables we can conclude that andre has heart disease.

here the heart disease is the sensitive attibute which we want to publish.

Attacks against k anonymity:

1. Complementary Release Attack

Different releases can be linked together to compromise k-anonymity.

Solution:

Consider all of the released tables before release the new one, and try to avoid linking. Other data holders may release some data that can be used in this kind of attack. Generally, this kind of attack is hard to be prohibited completely.

Example:

Race	BirthDate	Gender	ZIP	Problem
black	9/20/1965	male	02141	short of breath
black	2/14/1965	male	02141	chest pain
black	10/23/1965	female	02138	painful eye
black	8/24/1965	female	02138	wheezing
black	11/7/1964	female	02138	obesity
black	12/1/1964	female	02138	chest pain
white	10/23/1964	male	02138	short of breath
white	3/15/1965	female	02139	hypertension
white	8/13/1964	male	02139	obesity
white	5/5/1964	male	02139	fever
white	2/13/1967	male	02138	vomiting
white	3/21/1967	male	02138	back pain

Race	BirthDate	Gender	ZIP
black	1965	male	0214
black	1965	male	0214
black	1965	female	0213
black	1965	female	0213
black	1964	female	0213
black	1964	female	0213
white	1964	male	0213
white	1965	female	0213
white	1964	male	0213
white	1964	male	0213
white	1967	male	0213
white	1967	male	0213

2.K-Anonymity does not provide privacy if:Sensitive values in an equivalence class lack diversity

The attacker has background knowledge:

Zipcode	Age	Disease
476**	2*	Heart Disease
476**	2*	Heart Disease
476**	2*	Heart Disease
4790*	=40	Flu
4790*	=40	Heart Disease
4790*	=40	Cancer
476**	3*	Heart Disease
476**	3*	Cancer
476**	3*	Cancer

Homogenity attack:

Bob	
Zipcode	***Age***
47678	27

we can conclude that bob has heart disease.

Background knowledge Attack:

Carl	
Zipcode	***Age***
47673	36

we can guess from the above table that carl has either heart attack or cancer

2.1.3 l-diversity

Distinct l-diversity:

Each equivalence class has at least l well-represented sensitive values

Limitation:

Doesn't prevent the probabilistic inference attacks

Ex.
In one equivalent class, there are ten tuples. In the "Disease" area, one of them is "Cancer", one is "Heart Disease" and the remaining eight are "Flu". This satisfies 3-diversity, but the attacker can still affirm that the target person's disease is "Flu" with the accuracy of 70%.

Entropy l-diversity:

Each equivalence class not only must have enough different sensitive values, but also the different sensitive values must be distributed evenly enough.

In the formal language of statistic, it means the entropy of the distribution of sensitive values in each equivalence class is at least log(l)

Disadvantages of l -diversity

1 .Two sensitive values

HIV positive (1%) and HIV negative (99%) we conclude that nearly all the people have aids.

1. **Similarity attack:**

Bob	
Zipcode	***Age***
47678	27

Zipcode	Age	Salary	Disease
476**	2*	20K	Gastric Ulcer
476**	2*	30K	Gastritis
476**	2*	40K	Stomach Cancer
4790*	=40	50K	Gastritis
4790*	=40	100K	Flu
4790*	=40	70K	Bronchitis
476**	3*	60K	Bronchitis
476**	3*	80K	Pneumonia
476**	3*	90K	Stomach Cancer

from the above we can conclude that bob has stomach disease.
l-diversity does not deal with the meaning of the sensitive attributes.

2.2 Dynamic data privacy prevention methods:

2.2.1 Perturbation:

Pertubation means deviation of published data from the original data .In this we modify data in a way such that the modified data produces same result as that of original data this is done using mathematical functions.

Advantages:

privacy is preserved to a very high extent .

Limitations:

this can be used when data has fixed utility.
too much noise can destroy data quality.

2.2.2 Tree structure:

in this tuples are converted into tree nodes ,when the size of the tree reaches k the tuples are published and the tree is modified here k is the number of tuples to be published

Advantages :

its structured.

Disadvantages:

the information is high due to size limitation.

2.2.3. Adding l-diversity sensitive data:

instead of quasi identifier generalization l-diversified data streams are added to original data .

Advantages:

its difficult to breach privacy

Disadvantage:
its time consuming.

3.0 RELATED WORK

3.1 Faanst Algorithm:

parameters used in the algorithm k,u,D
k- defines the parameter for cluster anonymization
d-defines the number of clusters which can be used later
u-defines the processing window size

Algorithm:

When the numbers of tuples in the processing window reaches μ, one round of the clustering algorithm is started slide again in order to accumulate more tuples in each round.

Drawback:

The main drawback of FANNST is that some tuples may remain in the system more than allowable time constraint.

In addition, the time and space complexity of the algorithm is O(S*S) and not efficient for a data streaming algorithm . Another weakness of FANNST is that it does not support categorical data.

3.2 Fads Algorithm:

The algorithm considers a set as a buffer and saves at most δ tuples in it. Also, another set (setkc) is considered to hold the newly created cluster for later reuse.

Each k-Anonymized cluster will be remained in setkc up to the reuse constraint Tkc and after that the cluster is removed.

Drawbacks:

The main drawback of the FADS is that the algorithm does not check the remaining time of tuples that hold in the buffer in each round and are outputted them when they might be considered to have expired.

The other important weakness of FADS is that it is not parallel and cannot handle a large amount of data streams in tolerable time.

4.0 PARAMETERS OF ALGORITHM

Data Stream:

A sequence of tuples is defined as < sn >n∈N where N is the natural number set.the kth term of<sn>is orders pair (t,Tk) where k is a number and tk is tuple.

A data stream S is a potentially infinite sequence of tuples, depicted by < ti >, where all tuples ti follow the schema ti =< ID,a1,,am,q1,,qn,TS >. ID is an identifier attribute; q1 to qn are quasi identifiers and TS is the time stamp.

Cluster:

Cluster is a set of tuples in a stream. Suppose that PS is a set of tuples in stream Cluster C can be defined as follow:

C ={t|t belongsPs};

K-anonymized cluster :

If a cluster C built from data stream and the number of unique tuple in the cluster is

greater than k, the cluster is called a k-anonymized cluster.

Generalization:

Generalization is a function that maps a cluster into a tuple. More formally, generalization function G is defined as G : PowerSet(TUPLE) → TUPLE where TUPLE is the set of all possible tuples.

Numerical value generalization:

Numerical values are generalized in between maximum and minimum value i,e they are generalized in their domain.

Categorical value generalization:

categorical values are generalized to their lowest common ancestors.

eg of above two types of generalization:

considering a cluster of 3 tuples which contains both numerical and categorical values .

the tuples contain name ,profession and age of employees .

C=<“prof.young”, Academic ,43>

,<“Mr.Zhou”,non-Academic,39> ,

<“Prof.Chung” ,Academic,46 >.

the above tuple can be generalized as follows: gc =< ∗,staff,[39—46] >.since we don’t want to disclose the name we kept * in first column here profession is categorical value and age is numerical value age is generalized as [max,min] and profession is generalized to lowest common ancestor of academic and non academic.

Distance:

Distance is used to calculate the similarity or dissimilarity between two tuples .this function is the heart of the clustering. Generally clustering is done based on distance calculation the tuples with closest distance are placed the same cluster.

Types of distances:

1.Distance between the numerical values

let v1,v2 be 2 numerical values

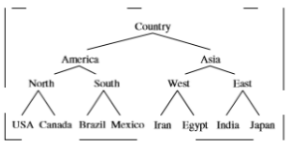
the distance between v1,v2 =d(v1,v2)=|v1-v2|/|D|

where D is the domain of the values

2. Distance between 2 categorical values :

If all the categorical values are arranged in the form of a tree where root is the most generalized value of all the values and lowest most level containing ore specialized values of the categorical values

e.g of a categorical tree



distance between two categorical values v1,v2 =d(v1,v2)=(height of the subtree rootes at lowest common ancestor of (v1,v2))/ (height of tree)

eg :distance between india and egypt(considering the tree from the above picture)

=height of subrooted tree of lowest common ancestor of (india,egypt)/height of the tree

=height of tree with east as root /height of tree=1/3=0.33

Distance between two tuples:

distance between two tuples t= {N1,...,Nm, C1,...,Cn} be the quasi-identifier of table T, whereNi(i =1,...,m) is an attribute with a

numeric domain and Cj(j =1,...,n) is an attribute
with a categorical domain.
The distance d(r1,r2) (I.e the distance between 2 tuples r1,r2) is defined as:
d(r1,r2)=**sum of distances between numerical attributes of two tuples +sum of distances between categorical attributes of two tuples.**

Information Loss: generalization leads to information loss, but we have to group clusters in such a way that the information loss is minimum.
information loss of a single cluster is caluculated as:
Total Information Loss=sum of information loss of all the clusters.
Information Loss Of The Cluster=info loss of all the tuples in the cluster.
Information Loss Of The Tuple= information loss of all the attributes
(categorical attributes and numerical attributes.)
Information Loss Of Numerical Attribute= (value of attribute) / (domain of the attribute).
Information Loss Of Categorical Attribute=(Height of the tree rooted with categorical attribute) / (Height of Categorical attribute tree).
Where h is the height of the tree and k is height of the tree rooted at the required categorical attribute.

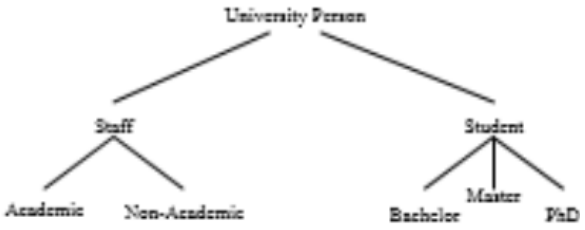
5.0 PROPOSED ALGORITHM

5.1 Details of the Algorithm:
S=Total number of tuples in the datset.
K=Anonymization parmeter.
$=number of tuples to be read before processing.
SetTp=Set of $ tuples.
SetKc=Set of all unique generalized set.
Snew=Set of K tuples.
Gs=Generalized set of Snew.

The algorithm reads $ tuples continuously and insert them into the SetTp. At First For each tuple in SetTp procedure finds t's K-1 nearest tuples in SetTp. With the help of tuple t and it's K-1 nearest tuples,generate a new set called as Snew and generalize it into Gs. Then a set with minimum information loss (Sk-best) that covers tuple t is chosen from SetKc If Sk-best exist and has smaller information loss compared to Gs then tuple t is published Sk-best generalization.

If tuple t does not match with any set in SetKc which has less information loss compared to Gs ,then tuple t is published with Snew generalization i.e Gs. Then Gs is inserted in SetKc .

In the following, a simple example is illustrated for better understanding. Assume that Table 1 is a portion of a university data stream, in which quasi-identifier are age and job. Also $ and K are assumed as $ = 3 and K = 2. Suppose that in thread n the value of variables are as follows:



In this stage, information loss of Sk—best is compared with Gs information loss. As, The information loss of Sk—best is less than Gs , tuple with idn is published with Sk—best generalization. The published tuples are shown in Table 2.

Pid	Age	University-Person
Id1	22	Bachelor
Id2	24	Master
Id3	37	Non-academic
. . .	. . .	. . .
Idn	45	Academic
Idn+1	26	Non-Academic
Idn+2	39	Phd

Table 1: University person

Pid	Age	University Person
Id1	[22-24]	Student
Id2	[22-24]	Student
Id3	[15-95]	Person-university
. . .	. . .	. . .
Idn	[44-46]	Staff
Idn+1	[26-39]	University-Pesrson
Idn+2	[26-39]	University-Person

Table2:Two Anonymized University-person

- SetTp = {(<idn,45,academic>, <idn+1,26,Non—academic>,<idn+2,39,PhD>) }
- SetKc={(([22—24],university), ([31—39],staff),([44—46],staff))}
- Snew=(<idn,45,academic>,<idn+2 ,26,non—academic>)
- Gs = ([26—45],staff)
- Sk—best = ([44—46],staff)

Algorithm:

Big data anonymization(S,K,$)

{

while S!=0 **do**

Read $ tuples and insert them into SetTp.

For each tuple t **do**

1. Select K-1 unique tuples which are closest to t among the tuples in SetTp and insert them into set Snew.

2. Generalize Snew into Gs.

For each set which covers t **do**

Calculate the information loss

End for

3. Select a set which incures less information loss
Call the set as Sk-best

4. **If** (Sk-best exist and Sk-best generate less information loss
Than Gs) **then**
Publish t with Sk-best generalization
Else
Publish t with Gs and insert Gs in set Kc
End if

End for

End while

}

6.0 RESULT AND DISCUSSION

Experiment environment:

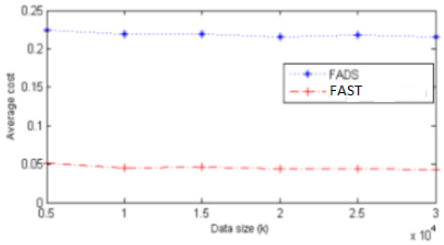
This experiment is performed on the system having intel i5processor with processing power of 2.2Ghz and main memory of 4.0 GB using linux platform .The algorithm is implemented in java and executed with the help of hadoop map reduce framework.

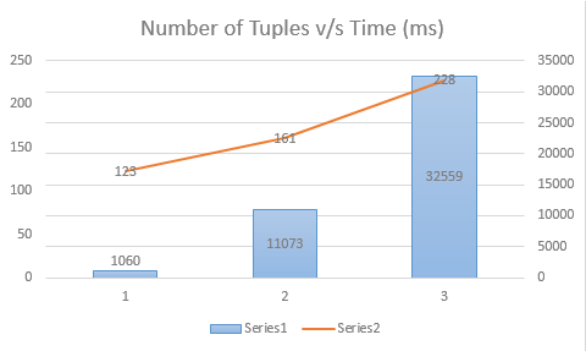
Dataset Description:

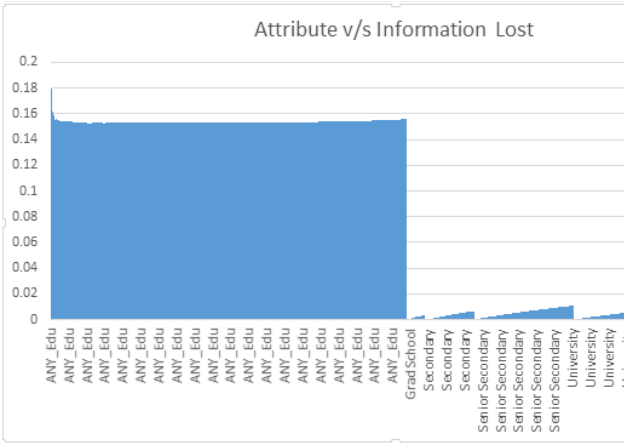
In this experiment we evaluated the performance of proposed algorithm on the Adult dataset from UCI[2].The data set was widely used in privacy preserving literature. The taxonomy tree is defined as per figure .The sensitive attribute in dataset is age(numerical) and profession(categorical).

Parameters and improvements :

The total number of records in the data set used for the experiment is 32,599 tuples. The algorithm efficiency is verified with parameters average information loss. The average information loss of the proposed algorithm,FADS and FAST is presented in the figure . The algorithm publishes data with less information loss, because of the SetKc in the proposed approach has more entities so that the data tuple has more options to select and this decreases the information loss and hence the algorithm improves.The average execution time drastically decreases as map-reduce is used.







FUTURE WORK

The methods used in this project are at a very rudimentary level. We have implemented the proposed algorithm in a single node cluster. Efficiency can be improved by adding more nodes to the Hadoop cluster

CONCLUSION

All the algorithms which are present for data stream processing are not capable of processing big data i.e data with high capacity and volume .The data which is processed using data anonymization (non-parallel) algorithms uses old languages (java, sql) and old techniques which are not very efficient i.e they take lot of time for computation and sometimes provide tuples which are expired, this lead to loss of accuracy and loss of privacy which is very dangerous. Static algorithms need all the computations to be performed on a single node due to which the data and the processing requirements are very high and the computers used are prone to failure which is very expensive to recover.

In this paper we have proposed an algorithm (big data anonymization) which uses hadoop frame work to process the data .Using hadoop the computers resources are used to maximum extent by which time required for computation is reduced which in turn prevents the publishing of expired tuples. Other advantages of this algorithm is that computations can be performed on nodes which have less computation and storage capacity compared to that of computers which perform non parallel data processing.

Using hadoop the failures in both data and processors can be recovered. This features drastically reduces the maintenance cost and the initial setup cost.

REFERENCES

[1] C. C. Aggarwal, J. Han, J. Wang, and P. S. Yu. A framework for clustering evolving data streams. In Proceedings of the 29th international conference on Very large data bases-Volume 29, pages 81–92. VLDB Endowment, 2003.

[2] C. Blake and C. J. Merz. {UCI} repository of machine learning databases. 1998.

[3] J. Cao, B. Carminati, E. Ferrari, and K.-L. Tan. Castle: Continuously anonymizing data streams. Dependable and Secure Computing, IEEE Transactions on, 8(3):337–352, 2011.
[4] C. Dwork. Differential privacy. In Automata, languages and programming, pages 1–12. Springer, 2006.
[5] B. Fung, K. Wang, R. Chen, and P. S. Yu. Privacy-preserving data publishing: A survey of recent developments. ACM Computing Surveys (CSUR), 42(4):14, 2010.
[6] B. C. Fung, K. Wang, and P. S. Yu. Top-down specialization for information and privacy preservation. In Data Engineering, 2005. ICDE 2005. Proceedings. 21st International Conference on, pages 205–216. IEEE, 2005.
[7] K. Guo and Q. Zhang. Fast clustering-based anonymization approaches with time constraints for data streams. Knowledge-Based Systems, 46:95–108, 2013.
[8] S. Kim, M. K. Sung, and Y. D. Chung. A framework to preserve the privacy of electronic health data streams. Journal of biomedical informatics, 2014. [9] F. Li, J. Sun, S. Papadimitriou, G. A. Mihaila, and I. Stanoi. Hiding in the crowd: Privacy preservation on evolving streams through correlation tracking. In ICDE, volume 1, page 2, 2007.
[10] N. Li, T. Li, and S. Venkatasubramanian. t-closeness: Privacy beyond k-anonymity and l-diversity. In ICDE, volume 7, pages 106–115, 2007.
[11] A. Machanavajjhala, D. Kifer, J. Gehrke, and M. Venkitasubramaniam. l-diversity: Privacy beyond k-anonymity. ACM Transactions on Knowledge Discovery from Data (TKDD), 1(1):3, 2007.