A Survey on Centralised and Distributed Approaches for
Subtree Anonymization

Bhushan Kakad\textsuperscript{a}\textsuperscript{*}, Jai Patil\textsuperscript{b}

\textsuperscript{a} PG Student at SSGMCE Shegaon, Department of Computer Science and Engineering, SGBAU Amaravati University.
\textsuperscript{b} Assistant Professor at SSGMCE Shegaon, Department of Computer Science and Engineering.

Abstract

Nowadays, data privacy is the most important task, mainly on large scale data set. Two approaches BUG and TDS are the two ways to do anonymization through classification as both are iterative processes. Both the approaches are good for certain value of k anonymity parameter, but not all the values of k. If we combine both the BUG and TDS, results are in the form of high gain scalability. So this hybrid approach is used to increase the efficiency as well as improve the scalability. But these are centralised approaches and suffers from the scalability problem. So TPTDS and MRTDS Framework are introduced for these purpose. In this paper we discuss the this popular centralised and distributed approaches for sub tree anonymization.

Keywords: Data anonymization; Bottom up generalization; Two phase top-down specialization; MapReduce, cloud; privacy preservation.

\section{Introduction}

Nowadays cloud services and the big data applications play the very important role in IT industry. For taking the advantage of cloud services big data applications prefer to go to cloud. Example for these are applications under in healthcare data and transactional data. Microsoft Health Vault an online cloud health service, aggregate data from users and shares the data from research institutes. So the big problem is to achieving the privacy of these data. Hence if one value of the sensitive attribute is leaked then it may cause the huge information loss on these services. So before releasing the data lots of privacy issues are likely to be urgently considered before the data sets are released or sheared on the cloud. It requires masking the sensitive data. Hiding the identity of an individual is nothing but data anonymization. There are various methods of achieving anonymity. Sub tree anonymization is the widely used method for data anonymization. Sub tree anonymization can be accomplish by two methods. One is the Bottom up generalization and another one is Top down specialization. Because of the
lacking of the parallelisation capability, both of the methods can't perform well for certain values of k. The hybrid approach[9] is used to increase the efficiency and the scalability of the sub tree anonymization

1.1. k-anonymity and l-diversity

k anonymity and l-diversity are the most widely used model for protecting privacy. k anonymised dataset has the property that each record is in distinguish by at least k-1 other records within the dataset[7]. The privacy is achieved highly if the value of k is large. Consider the same example as[10]. Now the person specific data is released by data provider. The quansi identifier consist of combination of attributes, eg {birthplace, birthyear, sex}. This information can't identify individual but the information can be used by others to link the person specific information and other related information. So individual may be identifiable. This privacy requirement is called as k-anonymity. For measuring the optimality of k anonymization two types of cost matrices are used. First one is based on minimal generalization [13][14] and second is suggested by R.Agrawal[7]. The second approach i.e optimal k anonymization is not suitable because masking the noises and structures gives different effects. Another method for anonymity is l-diversity. l-diversity corresponds to some notion of linking quansi identifier QID with some other particular sensitive value[11]. l-diversity principle states that sensitive attribute would have at most same frequency, and l-diversity is the requirement that values of the sensitive attribute are well represented in each group.[5][6]. If data provider wants to discover intruder then he can do this with high probability distribution of the attributes. The disadvantage of this method is homogeneity and background knowledge attack has lacked. Further methods are introduced that will give more clarity regarding anonymization.

2. Bottom Up Generalization Approach

Let the data holder wants to release the person specific data R(S1,...Sn, cls) to the public. a record has the form <v1,...vn, cls>.Consider R records is sharing some attributes with external data source E. The value R∩E gives more specific value, So the probability of getting the real life value is high[3].

2.1 Anonymity

Virtual identifier VID is nothing but attributes share by R and E. Number of records in R with value vid on VID denotes a(vid). The minimum a(vid) for any value. The minimum a(vid) for any value vid on VID denotes A(VID). Anonymity requirement <VID,k> is satisfied by R iff A(VID)≥k. The value of K is specified by data holder

2.2 Generalization

Generalization is denoted by {c} → p. The meaning of the sentence is replacement of of child value c with the parent value p. Author K. Wang said the generalization is valid if and only if all the values below p are generalised to c.

2.3 Metrics for generalization

k-anonymity and preservation of information for classification should be achieved by good generalization. Generalization G: {c} → p., Rc is the set of records containing c and Rp denotes the set of records containing p.

\[ R_p = \sum_c |R_c| \]

The result of generalization G is nothing but the information loss and anonymity gain after replacing Rc with Rp. K Wang. adapt the entropy based on information loss and it can be substituted by other information measure.

\[ I(G) = \text{Inf}o(R_p) - \sum_c \frac{|R_c|}{|R_p|} \ast \text{Inf}o(R_c) \]  

(1)

\[ \text{Inf}o(Rx) \] is the entropy of Rx[8].
\[ Inf o(Rx) = - \sum_{cls} \frac{freq(Rx,cls)}{Rx} \log_2 \frac{freq(Rx,cls)}{Rx} \]  \hspace{1cm} (2)

\( freq(Rx,cls) \) is nothing but the number of records in the \( Rx \) with the class label \( cls \).

\( A(VID) \) and \( A_G(VID) \) denotes the anonymity before and after applying \( G \). \( A_G(VID) - A(VID) \) represent the anonymity gain.

If the value of \( k \) is so large then data holder have to specify larger \( k \) in first place. The modified anonymity gain is \( P(G) = x - A(VID) \)

\[ x = A_G(VID) \quad \text{if} \quad A_G(VID) \leq k \]

\[ x = k \quad \text{Otherwise}. \]

Now to reduce the information loss for each unit of anonymity gain for the given value of \( k \)

\[ \text{Minimize : } \frac{IP(G)}{P(G)} \]

Algorithm BUG

1. while \( R \) does not satisfy given anonymity requirement.
2. For all generalization \( G \) do
3. compute \( IP(G) \)
4. end for.
5. find the best generalization \( G_{best} \).
6. generalize \( R \) by best generalization.
7. end while
8. output \( R \)

Our generalization \( \{ c \} \rightarrow p \) value of \( | Rc | \) and \( freq(Rc,cls) \) are updated at each iteration. \( I(G) \) can be easily computed by the values of \( | Rc | \) and \( freq(Rc,cls) \).

3. Top Down Specialization

Top down specialization is an iterative process. TDS generalize the table by specializing it iteratively starting from most general state.[4]. In TDS parent value in the tree is specialized into child values.TDS is used for anonymize both the categorical as well as continuous attribute.

3.1 Specialization

The specialization \( v \rightarrow \text{child}(v) \), means that replacement of the parent \( v \) with child value in \( \text{child}(v) \) that generalizes the domain value in record. The function of specialization is to increase the information gain and decrease the anonymity because the records are more distinguishable by specific values. After applying the sequence of specialization starting from most general state in which each attribute has the top most value of its taxonomy tree \( T \) can be generalised[4]. Specialization gives us a maximum information gain and represented as \( \text{InfoGain}(v) \), and anonymity loss is denoted as \( \text{AnonyLoss}(v) \). Fung told that the specialization on \( v \) that has the maximum information gain for each unit of anonymity loss.
\[
\text{score}(v) = \frac{\text{InfoGain}(v)}{\text{AnonyLoss}(v)}
\]

\[
\text{InfoGain}(v) = I(R_v) - \sum_c \frac{|R_c|}{|R_v|} \cdot R_c
\]

\[
I(R_x) = \text{entropy of } R_x
\]

\[
I(R_x) = - \sum_{\text{cls}} \frac{f_{\text{freq}(R_x, \text{cls})}}{|R_x|} \cdot \log_2 \frac{f_{\text{freq}(R_x, \text{cls})}}{|R_x|}
\]

And

\[
\text{AnonyLoss}(v) = \text{avg}\{A(VID_i) - A_v(VID)\}
\]

### 3.2 Solution Cut

Subset of values in the tree that consist of only one value from root to its leaf of the taxonomy tree \( T \) called as solution cut. The purpose of the solution cut is nothing but preservation of the maximum information for classification

Algorithm TDS

1. Initialization of each value in \( T \) to top most value.

2. Initialize \( \text{CUT}_i \)

3. While \( r \in \text{CUT}_i \) is valid and beneficial do

   3.1 Search the best specialization from \( \text{CUT}_i \)

   3.2 Performing Best on \( T \) and update \( \text{CUT}_i \)

   3.3 Updating score and validity for \( r \) and \( r \in \text{CUT}_i \)

4. End while

5. Return Generalised \( T \) and \( \text{CUT}_i \)

### 3.3 Find The Best Specialization

IGPL is a search metric that is used to measure the correctness of specialization process. For doing this task it needs all the nodes of TIPS[4]. This step make the use of computed \( \text{InfoGain}(x) \) and \( A_r(VID_i) \) for all candidate \( r \) in \( \text{CUT}_i \), and computed \( A(VID_j) \) for each \( VID_j \). We already computed this information at starting of each iteration. Finding the Best consist of at most \( |\text{CUT}_i| \). Computation of score without accessing data records and also it consist of updation of the InfoGain \( r \). And \( A_r(VID_j) \).

### 4. Two phase Top Down Specialization

Broadly TPTDS is categorised as three main steps data partition, anonymization level merging and data specialization.
4.1 Problem Analysis

TDS gives the scalability problem when data set sufficiently large. The centralised TDS approaches in [12][10]. Centralised approaches fail due to the updation of statistic information and linkage structure. Distributed TDS approach [2] is proposed to addressed the distributed anonymisation problem which mainly concern privacy protection against other parties rather than scalability issue and again issues due to communication protocol and fault tolerance needs to be considered. Hence it is not possible for distributed algorithm to produce anonymization dataset.

4.2 TPTDS Approach

Two Phase Top Down Specialization Approach is based on the concept of parallelization. Two types of parallelization are exist in the Big data world. First one is job level and another one is task level, for making the whole use of cloud infrastructure. e.g Amazon elastic MapReduce service is the job level parallelism.

Algorithm TPTDS
Input: Data set D, Anonymity parameter k, kI , and Number of partitions P.
Output: Anonymized Data set D.

1. Partition S into Si
2. Execute $\text{MRTDS}(S_i, k^I, AL^D_1) \rightarrow AL^I_i, 1 \leq i \leq p$ in parallel as multiple MapReduce jobs.
3. Merge all intermediate anonymization levels into one, $\text{merge}(AL^I_1, AL^I_2, ..., AL^I_p) \rightarrow AL^I$.
4. Execute $\text{MRTDS}(S, k, AL^I) \rightarrow AL^*$ to achieve k-anonymity.
5. Specialize S according to $AL^*$, Output $S^*$.

4.3 Data Partition

Dividing large data records are done through random sampling technique is used. For each data record random number rand $1 \leq \text{rand} \leq p$ is generated. One thing kept in mind that for number of reducer should be equal to P. So each reduce handle only one value of rand exactly producing P resultant files.

4.4 Anonymization Level Merging

Next step is to merging all the anonymization levels into one. The process of merging of anonymization level is completed by merging cuts. The intermediate anonymisation level $k^I$-anonymity. The merged intermediate anonymization level AL would satisfy $k^I$-anonymity where $AL \leftarrow \text{MERGE}(\langle AL^I_1, AL^I_2, ..., AL^I_p \rangle)$ $k^I \geq k^I$.

4.5 Data specialization

For anonymization purpose on the original dataset S is specialised in a MapReduce job. In this Map function gives anonymization. In this Map function gives anonymization records and its count and Reduce function simply aggregates anonymous records and counts their number.

Algorithm
Data Specialization Map and Reduce
Input: Data record, Anonymization level $AL^*$. 
Output: Anonymous record.

Map: Construct anonymous record using partition and sensitive value.

Reduce: Emit sum.

4.6 MapReduce Version Of Centralized TDS

Usually, it is not possible for single MapReduce job to perform in a driver program to achieve such objective. The process consist of drivers of MRTDS and the two types of job. IGPL initialization and IGPL update. MRTDS driver coordinates the Map and Reduce jobs of the process.

4.7 IGPL Initialisation job

MRTDS produces the same anonymous data as the centralised TDS except the calculating the IGPL values [1]. As the name suggest IGPL initialisation job initialise information gain and privacy loss of all specialization in initial anonymisation level AL.

4.8 IGPL Update Job

IGPL update job directly affect on scalability and efficiency of MRTDS as it is executed iteratively. Iterative MapReduce jobs have not been supported by standard MapReduce framework like Hadoop[1]. Both the jobs IGPL initialisation and IGPL update are same but IGPL update job requires less computation and it consumes less network bandwidth.

Fig 1:- Execution Framework of MRDTS

The above fig shows that the framework based on standard MapReduce for the explanation of how datasets are being processes in MRTDS. Solid arrow indicates that data flows in canonical MapReduce framework. For handling iterations, curve arrow show the data flow. AL driver controlled the iterations of the MapReduce. AL is separated from the driver to all the workers including Mappers and Reducers via distributed cache mechanism. According to output of the IGPL Initialization or IGPL Update job, the value of the AL varies. For reducing the communication traffic combiner mechanism is allowed by MRTDS which will help to aggregates key value pair with the same key into one on the nodes running map functions. For further reducing the traffic MD5 is employed.
to compress the records transmitted for anonymity classification.

5 Performance Analysis
Following table help us to understand the advantage and Limitations of the above discuss methods.

Table 1. Overview of Centralised and Distributed Anonymisation Methods.

<table>
<thead>
<tr>
<th>Type</th>
<th>Method</th>
<th>Advantage</th>
<th>Limitations</th>
</tr>
</thead>
<tbody>
<tr>
<td>Centralised</td>
<td>Bottom Up Generalization</td>
<td>Transformation of Specific Data to Less Specific</td>
<td>Handles only categorical data</td>
</tr>
<tr>
<td>Centralised</td>
<td>Top Down Specialization</td>
<td>Handle both categorical as well as continuous data</td>
<td>Scalability issue when dataset is large</td>
</tr>
<tr>
<td>Distributed</td>
<td>Two phase top down specialization</td>
<td>Pararallisation and Scalability problem is solved</td>
<td>Communication traffic is high.</td>
</tr>
<tr>
<td>Distributed</td>
<td>MRTDS Framework</td>
<td>Reduce communication traffic.</td>
<td>Transmission overhead is caused due to data splitting.</td>
</tr>
</tbody>
</table>

6 Conclusion
In this paper we discuss the various anonymization technique for preserving the privacy of large collection of information. Each method of anonymization has some advantage as large dataset. TPTDS significantly improves the scalability and efficiency TDS over existing TDS approaches. MRTDS process helps us to understand the computation of IGPL and execution of Map and Reduce functions. BUG is used only for categorical attribute. TDS serves the approach that are used on both categorical and continuous attribute.

References
1. Xuyun zang, Laurence T. Chang Liu, Jinjun Chan "scalable Two Phase Top Down Specialization approaches for data anonymization".
2. Centralised and distributed anonymization for high dimensional healthcare data by N. Mohammed.
4. B. Fung, K. Wang, and P. Yu. Top-down specialization for information and privacy preservation. In ICDE, 2005
6. Yan zhao, Ming du and jiajin Le Survey on privacy preserving approaches in data publishing. In First international workshop on Database Technology and Applications in 2009.
9. Xuyun zang, chang Lue, Surya Nepal. Combining Top-Down and Bottom-Up: Scalable Sub-Tree Anonymization.


