A Framework for Privacy-Preserving Cluster Analysis

Benjamin C. M. Fung
CIISE, Concordia University
Montreal, QC, Canada H3G 1M8
Email: fung@ciise.concordia.ca

Ke Wang
Simon Fraser University
Burnaby, BC, Canada V5A 1S6
Email: wangk@cs.sfu.ca

Lingyu Wang Mourad Debbabi
CIISE, Concordia University
Montreal, QC, Canada H3G 1M8
Email: {wang, debbabi}@ciise.concordia.ca

***Abstract*—Releasing person-specific data could potentially reveal sensitive information of individuals. k-anonymization is a promising privacy protection mechanism in data publishing. Though substantial research has been conducted on k-anonymization and its extensions in recent years, few of them consider releasing data for a specific purpose of data analysis. This paper presents a practical data publishing framework for determining a generalized version of data that preserves both individual privacy and information usefulness for cluster analysis. Experiments on real-life data suggest that, by focusing on preserving cluster structure in the generalization process, the cluster quality is significantly better than the cluster quality on the generalized data without such focus. The major challenge of generalizing data for cluster analysis is the lack of class labels that could be used to guide the generalization process. Our approach converts the problem into the counterpart problem for classification analysis where class labels encode the cluster structure in the data, and presents a framework to evaluate the cluster quality on the generalized data.**

I. Introduction

After the series of worldwide terrorist attacks, law enforcement agencies have received more pervasive authorities to counter security challenges [18]. In September 2004, the Department of Homeland Security in the United States awarded $9 million research grants to foster and evaluate uses of "state-of-the-market" information technology that will improve information sharing and integration among the network of security agencies [19]. Recent research [4], however, indicates that the public feels an increased sense of intrusion and loss of privacy due to the increasing scope of information sharing among agencies, corporations, and governments. In a broad sense, there is a demand for simultaneous information sharing and privacy protection. This paper presents a technical response to the demand for the task of cluster analysis.

Consider a person-specific data table T about patients' information on *Zip code*, *Birthplace*, *Sex*, and *Disease*. The data holder wants to publish T to some recipient for cluster analysis. However, if a description on (*Zip code*, *Birthplace*, *Sex*) is so specific that few people match it, publishing the table will lead to linking a unique or a small number of individuals with the sensitive information on *Disease*. Even if the currently published table T may not contain sensitive information, individuals in T can be linked to the sensitive information in some external source by a join on the common attributes [17]. In this paper, we want to determine a generalized version of T that satisfies two requirements: the *privacy requirement* and the *clustering requirement*.

Privacy Requirement: To protect privacy, instead of publishing the raw table T, the data holder publishes a generalized table $T(QID, Sensitive\_attribute)$, where QID contains a set of generalized identifying attributes, such as, (*Zip code*, *Birth Region*, *Sex*). By replacing *Birthplace* with *Birth Region*, more records will match the generalized description, and therefore, individuals who match the description will become less identifiable. The privacy requirement is specified by k-anonymity [15][17]: Each record in table T shares the same value on QID with at least $k-1$ other records, where k is an anonymity threshold specified by the data holder.[1] All records in the same QID group are made indistinguishable, and therefore, difficult to determine whether a matched individual actually has the disease from T.

Clustering Requirement: We consider publishing a table T to a recipient for the purpose of cluster analysis where the goal is to group similar objects into the same cluster and group dissimilar objects into different clusters. We assume the $Sensitive\_attribute$ is important for the task of cluster analysis; otherwise, it should be removed. The recipient may or may not be known at the time of data publication.

We define the *anonymity problem for cluster analysis* as follows: *For a given anonymity requirement and a raw data table T, we want to determine an anonymous version of T that preserves as much as possible the information for cluster analysis.* There are many possible k-anonymous versions of T. The major challenge is how to pick the "appropriate" one for cluster analysis. An inappropriately generalized version could put originally dissimilar objects into the same cluster, or put originally similar objects into different clusters because other generalized objects become more similar to each other. Therefore, a quality-guided generalization process is crucial. Unlike the anonymity problem for classification analysis [6][7], the anonymity problem for cluster analysis does not have class labels to guide the generalization. It is not even clear what "information for cluster analysis" means and how to evaluate the quality of generalized data in terms of cluster analysis.

These challenges bring out the contributions of this paper:

[1] To avoid confusion with the variable k in k-means clustering algorithm, we use h to denote the anonymity threshold in the rest of this paper.

ISI 2008, June 17-20, 2008, Taipei, Taiwan
1-4244-2415-3/08/$20.00 ©2008 IEEE

(1) We define the anonymity problem for cluster analysis. (2) We present a framework to convert the problem to the counterpart problem for classification analysis. The idea is to extract the cluster structure from the raw data, encode it in the form of class labels, and preserve such class labels while generalizing the data. (3) Our proposed framework can address the unknown choices of the prospective recipient and permit the data holder to evaluate the cluster quality of the generalized data. (4) We experimentally study the effectiveness of the approach on real-life data. The results suggest that, by focusing on preserving cluster structure in the generalization process, the cluster quality is significantly better than the cluster quality on the generalized data without such focus. (5) We show the extensions to achieve other privacy notions [13][24][25].

Given that the clustering task is known in advance, why not publish the analysis result instead of the data records? Unlike classification trees and association rules, publishing the cluster statistics (e.g., cluster centers, together with their size, and radius) usually cannot fulfil the requirement of cluster analysis. Often, data recipients want to browse into the clustered records to gain more knowledge. For example, a police officer may browse into some clusters of criminals and examine their common characteristics. Thus, publishing data records often is a vital requirement for cluster analysis.

II. Related Works

Recently, many generalization methods [5][11][15] have been proposed to achieve k-anonymity [17]. Their work does not consider cluster analysis or any other specific use of data, but uses simple quality measures to guide generalization. Preserving anonymity for classification analysis was studied in [3][6][7][9][12][22][23]. The idea is using the available class labels to guide the generalization process so that the class labels can still be identified in generalized QID. In the case of cluster analysis, no class label is available for this purpose. Alternative privacy notions, such as ℓ-diversity [13] and confidence bounding [24][25], were proposed to hide the correlation between QID and sensitive attributes.

There is a family of anonymization methods [1][2] that achieves privacy *by* clustering similar data records together. Their objective is very different from our studied problem that is to publish data *for* cluster analysis. An anonymization approach, called *condensation* [1], proposes to first condense the records into multiple non-overlapping groups in which each group has a size of at least h records. For each group, extract some statistical information, such as sum and covariance, that suffices to preserve the mean and correlations across different attributes. Then, based on the statistical information, generate data records for each group. In a similar spirit, *r-gather clustering* [2] partitions records into clusters and releases the cluster centers, together with their size, radius, and a set of associated sensitive values.

Some secure protocols [8][20] have been proposed to determine a clustering solution from vertically and horizontally partitioned data owned by multiple parties. [26] studies the tradeoff between privacy protection and communication complexity of secure protocols for information sharing. In their models, accessing data held by other parties is prohibited. In contrast, our goal is to share generalized data.

TABLE I
THE LABELLED TABLE

Rec ID	Education	Sex	Age	...	Class	Count
1-3	9th	M	30		$0C_1$ $3C_2$	3
4-7	10th	M	32		$0C_1$ $4C_2$	4
8-12	11th	M	35		$2C_1$ $3C_2$	5
13-16	12th	F	37		$3C_1$ $1C_2$	4
17-22	Bachelors	F	42		$4C_1$ $2C_2$	6
23-26	Bachelors	F	44		$4C_1$ $0C_2$	4
27-30	Masters	M	44		$4C_1$ $0C_2$	4
31-33	Masters	F	44		$3C_1$ $0C_2$	3
34	Doctorate	F	44		$1C_1$ $0C_2$	1
				Total:	$21C_1$ $13C_2$	34

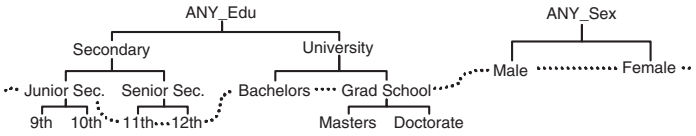


Fig. 1. A solution cut for $QID_1 = \{Education, Sex\}$

III. Problem Statements

A *labelled* table has the form $T(D_1, \ldots, D_m, Class)$ that contains a set of records of the form $\langle v_1, \ldots, v_m, cls\rangle$, where v_j is a domain value of attribute D_j and cls is a class label of the $Class$ attribute. Each D_j is either a categorical or a continuous attribute. An *unlabelled* table has the same form as a labelled table but without the $Class$ attribute.

Suppose that a data holder wants to publish a person-specific table T, but also wants to protect against linking an individual to sensitive information either inside or outside T through some sets of identifying attributes, called *quasi-identifiers*. A *sensitive linking* occurs if some value on a quasi-identifier is shared by only a "small" number of records in T. This requirement is formally defined below.

Definition 3.1 (Anonymity requirement): Consider p quasi-identifiers $QID_1, \ldots, QID_p$ on T, where $QID_i \subseteq \{D_1, \ldots, D_m\}$ for $1 \leq i \leq p$. $a(qid_i)$ denotes the number of data records in T that share the value qid_i on QID_i. The *anonymity* of QID_i, denoted by $A(QID_i)$, is the smallest $a(qid_i)$ for any value qid_i on QID_i. A table T satisfies the *anonymity requirement* $\{\langle QID_1, h_1\rangle, \ldots, \langle QID_p, h_p\rangle\}$ if $A(QID_i) \geq h_i$ for $1 \leq i \leq p$. QID_i and the *anonymity threshold* h_i's are specified by the data holder. If $QID_j \subseteq QID_i$ where $j \neq i$, then QID_j can be removed. ■

To achieve an anonymity requirement, we generalize attributes in $\cup QID_i$, for $1 \leq i \leq p$, on T according to some taxonomy trees. We assume that a *taxonomy tree* is specified by the data holder for each categorical attribute D_j in $\cup QID_i$. A leaf node represents a domain value and a parent node represents a less specific value. For a continuous attribute D_j in $\cup QID_i$, the taxonomy tree is grown at runtime, where each

node represents an interval, and each non-leaf node has two child nodes representing some "optimal" binary split of the parent interval [14].

Example 3.1: Consider the data in Table I and taxonomy trees in Figure 1. The table has 34 records, with each row representing one or more raw records that agree on (*Education*, *Sex*, *Age*). The *Class* column stores a count for each class label. The anonymity requirement $\langle QID_1 = \{Education, Sex\}, 4\rangle$ states that every existing qid_1 in T must be shared by at least 4 records in T. Therefore, $\langle 9th, M\rangle$, $\langle Masters, F\rangle$, $\langle Doctorate, F\rangle$ violate this requirement. To make the "female doctor" less unique, we can generalize *Masters* and *Doctorate* to *Grad School*. As a result, "she" becomes less identifiable by being one of the four females who have graduate degree in T. ■

A generalized T can be viewed as a "cut" through the taxonomy tree of each attribute in $\cup QID_i$. A *cut* of a tree is a subset of values in the tree that contains exactly one value on every root-to-leaf path. A *solution cut* is $\cup Cut_j$, where Cut_j is a cut of the taxonomy tree of an attribute D_j in $\cup QID_i$, such that the generalized T represented by $\cup Cut_j$ satisfies the anonymity requirement. Figure 1 shows a solution cut, indicated by the dash line, for the anonymity requirement in Example 3.1. The solution cut represents a maximally specialized table. Any further specialization on *Junior Sec.* or *Grad School* would violate the anonymity requirement.

Definition 3.2 (Anonymity problem for classification analysis): Given a labelled table T, an anonymity requirement $\{\langle QID_1, h_1\rangle, \ldots, \langle QID_p, h_p\rangle\}$, and a taxonomy tree for each categorical attribute in $\cup QID_i$, the *anonymity problem for classification analysis* is to generalize T on the attributes $\cup QID_i$ to satisfy the anonymity requirement while preserving as much as possible the information for classification analysis. ■

For classification analysis, the information utility of attributes can be measured by their power of discriminating class labels [3][6][7][9][12][22][23]. For cluster analysis, however, no class labels are available. What kind of information should be preserved for cluster analysis? One natural approach is to preserve the cluster structure in the raw data. Any loss of structure due to the anonymization is measured relatively to such "raw cluster structure." In this paper, we define the anonymity problem for cluster analysis as follows to reflect this natural choice of approach.

Definition 3.3 (Anonymity problem for cluster analysis): Given an unlabelled table T, an anonymity requirement $\{\langle QID_1, h_1\rangle, \ldots, \langle QID_p, h_p\rangle\}$, and a taxonomy tree for each categorical attribute in $\cup QID_i$, the *anonymity problem for cluster analysis* is to generalize T on the attributes $\cup QID_i$ such that the generalized table T satisfies the anonymity requirement and has a cluster structure as similar as possible to the cluster structure in the raw table T. ■

Intuitively, two cluster structures are similar whenever two objects belong to the same cluster, or different clusters, in one cluster structure, so do they in the other cluster structure. A formal measure for the similarity of two structures will be discussed in Section IV-C.

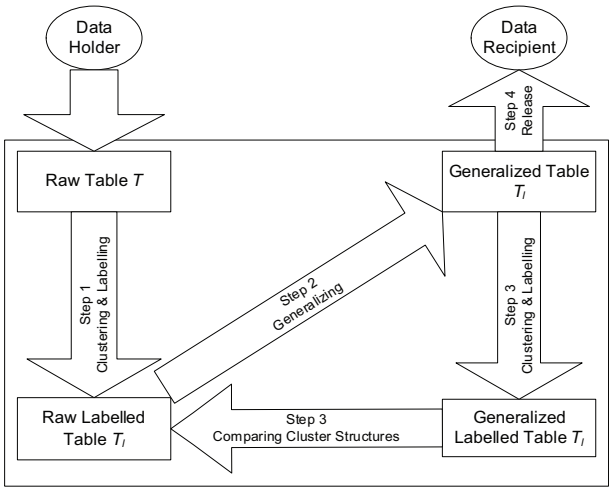


Fig. 2. The framework

IV. Our Approach

A. Overview of Solution Framework

First, we determine the raw cluster structure in the raw table T and label each record in T by a class label. This labelled table, denoted by T_l, has a $Class$ attribute that contains a class label for each record. Essentially, preserving the raw cluster structure is to preserve the power of discriminating such class labels during generalization. Generalization that diminishes the difference among records belonging to different clusters (classes) is penalized. As the requirement is the same as the anonymity problem for classification analysis, we can apply existing anonymization algorithms [3][6][7][9][12] to the anonymity problem for cluster analysis. Figure 2 summarizes the approach. We explain each step as follows.

Step 1: Convert T to a labelled table T_l. Apply a clustering algorithm to T to find the raw cluster structure, and label each record in T by its class label. The resulting labelled table T_l has a $Class$ attribute containing the labels.

Step 2: Generalize T_l. Employ an anonymization algorithm for classification analysis to generalize T_l to satisfy the given anonymity requirement.

Step 3: Evaluate the generalized T_l. Apply a clustering algorithm to the generalized T_l with the labels removed, where the number of clusters is the same as in Step 1, and compute the similarity between the cluster structure found and the raw cluster structure in Step 1. The similarity measures the degree of loss in cluster structure due to generalization. If the evaluation is unsatisfactory, the data holder may repeat Steps 1-3 with different number of clusters and choice of clustering algorithms. By default, the clustering algorithm in this step is the same as in Step 1, but can be replaced with the recipient's choice if this information is available. See more discussion below.

Step 4: Release the generalized T_l. If the evaluation in Step 3 is satisfactory, the data holder can release the generalized T_l together with some optional supplementary information: all the taxonomy trees (including those generated at runtime

for continuous attributes), the final solution cut, the similarity score computed in Step 3, and the $Class$ attribute in Step 1.

In some data publishing scenarios, such as data release from census bureau, the data holder does not even know who the prospective recipient is; therefore, does not know how the recipient will cluster the published data. How should the data holder set the cluster number? In this case, we suggest releasing one version for each reasonable cluster number so that the recipient can make the choice based on her desired number of clusters. Our previous work [5] presented an anonymization method to address the potential privacy breach caused by releasing different generalized versions of the same underlying data. Though [5] does not aim at preserving cluster structure, its quality-guided function can be modified to bias on preserving cluster structure encoded by the class labels proposed in this framework.

In the rest of this section, we describe the anonymization in Step 2 and the evaluation in Step 4, and their extensions.

B. Anonymization for Classification

Our implementation uses the top-down specialization (or TDS) in [6][7] due to its capability of anonymizing both categorical attributes and continuous attributes efficiently. In contrast, most bottom-up generalization methods [11] cannot generalize continuous attributes without taxonomy trees. TDS takes a labelled table and an anonymity requirement as inputs. The main idea of TDS is performing those generalizations that preserve the information for discriminating the class labels in a top-down manner. The next example illustrates this point.

Example 4.1: Suppose that the raw cluster structure produced by Step 1 has the class (cluster) labels given in the $Class$ attribute in Table I. In Example 3.1, we generalize *Masters* and *Doctorate* into *Grad School* to make linking through (*Education*,*Sex*) more difficult. No information is lost in this generalization because the class label C_1 does not depend on the distinction of *Masters* and *Doctorate*. However, further generalizing *Bachelors* and *Grad School* to *University* makes it harder to separate the two class labels involved. ■

Instead of generalizing a labelled table T_l starting from most specific domain values, TDS *specializes* it starting from the most general state where each categorical attribute has the top most value and each continuous attribute has the single interval covering all values. A specialization, written $v \rightarrow child(v)$ where $child(v)$ is the set of child values of v, refines the parent value v into a child value in $child(v)$ in all records containing v. The refinement must be consistent with the domain value in the raw record. For example, if the domain value is *Doctorate*, refining *University* into *Grad School* is consistent, but refining *University* into *Bachelors* is not. The top-down specialization is performed as follows.

First, all attributes not in $\cup QID_i$ are removed from T_l, and duplicates are collapsed into a single row with the $Class$ column storing the count for each class label. Initially, the solution cut $\cup Cut_j$ contains the top most value or interval for each attribute in $\cup QID_i$, and all records in T_l are generalized to such top most values. In each iteration, TDS pushes down the solution cut by specializing some value v in it. The specialization process stops if any further specialization will lead to a violation of the anonymity requirement.

TABLE II
THE ANONYMOUS TABLE WRT QID_1 AND QID_2

Rec ID	Education	Sex	Age	$\cdots$	Class	Count
1-7	Junior Sec.	ANY	[1-37)	$\cdots$	K_1	7
8-12	11th	ANY	[1-37)	$\cdots$	K_1	5
13-16	12th	ANY	[37-99)	$\cdots$	K_2	4
17-26	Bachelors	ANY	[37-99)	$\cdots$	K_2	10
27-34	Grad School	ANY	[37-99)	$\cdots$	K_2	8

TABLE III
THE SIMILARITY OF TWO CLUSTER STRUCTURES

Clusters in Table 1	Clusters in Table 3	
	K_2	K_1
C_1	19	2
C_2	3	10

Each specialization increases information utility and decreases anonymity because records are more distinguishable by specific values. At each iteration, TDS greedily selects the specialization that has the highest score, in terms of the information gain per unit of anonymity loss. We omit the detailed definitions that can be found in [6][7].

Example 4.2: Consider the labelled table in Table I and the anonymity requirement:

$\{\langle QID_1 = \{Education, Sex\}, 4\rangle, \langle QID_2 = \{Sex, Age\}, 11\rangle\}$.

Initially, all data records are generalized to ⟨*ANY_Edu*, *ANY_Sex*, *[1-99)*⟩ and $\cup Cut_j$ = {*ANY_Edu*, *ANY_Sex*, *[1-99)*}. To find the next specialization, we compute the score for each of *ANY_Edu*, *ANY_Sex*, and *[1-99)*. Table II shows the generalized data that satisfies the anonymity requirement after performing the following specializations in order:

[1-99) → {*[1-37)*, *[37-99)*}
ANY_Edu → {*Secondary*, *University*}
Secondary → {*Junior Sec.*, *Senior Sec.*}
Senior Sec. → {*11th*, *12th*}
University → {*Bachelors*, *Grad School*}. ■

C. Evaluation

This step compares the raw cluster structure found in Step 1 in Section IV-A, denoted by $\mathcal{C}$, with the cluster structure found in the generalized data in Step 3, denoted by $\mathcal{C}_g$. Because $\mathcal{C}$ and $\mathcal{C}_g$ are extracted from the same set of records, we can evaluate their similarity for the record groupings. We use the well-known *F-measure* [21] for measuring the similarity. The idea is to treat each cluster in $\mathcal{C}$ as the relevant set of records for a query, and treat each cluster in $\mathcal{C}_g$ as the result of a query. The clusters in $\mathcal{C}$ are called "natural clusters," and those in $\mathcal{C}_g$ are called "query clusters."

For a natural cluster C_i in $\mathcal{C}$ and a query cluster K_j in $\mathcal{C}_g$, let $|C_i|$ and $|K_j|$ denote the number of records in C_i and K_j, respectively, let n_{ij} denote the number of records contained in both C_i and K_j, and let $|T|$ denote the total number of records in T. The *recall*, *precision*, and *F-measure* for C_i and K_j are calculated as follows: $Recall(C_i, K_j) = \frac{n_{ij}}{|C_i|}$ is the fraction of relevant records retrieved by the query. $Precision(C_i, K_j) = \frac{n_{ij}}{|K_j|}$ is the fraction of relevant records

among the records retrieved by the query. $F(C_i, K_j)$ measures the quality of query cluster K_j in describing the natural cluster C_i, by the harmonic mean of $Recall$ and $Precision$.

$$F(C_i, K_j) = \frac{2 * Recall(C_i, K_j) * Precision(C_i, K_j)}{Recall(C_i, K_j) + Precision(C_i, K_j)} \quad (1)$$

The success of preserving a natural cluster C_i is measured by the "best" query cluster K_j for C_i. We measure the quality of $\mathcal{C}_g$ using the weighted sum of such maximum F-measures for all natural clusters. This measure is called the *overall F-measure* of $\mathcal{C}_g$, denoted by $F(\mathcal{C}_g)$:

$$F(\mathcal{C}_g) = \sum_{C_i \in \mathcal{C}} \frac{|C_i|}{|T|} max_{K_j \in \mathcal{C}_g}\{F(C_i, K_j)\} \quad (2)$$

Note that $F(\mathcal{C}_g)$ is in the range [0,1]. A larger value indicates a higher similarity between the two cluster structures generated from the raw data and the generalized data, respectively.

Example 4.3: The $Class$ column in Table II shows a cluster structure with cluster number $k = 2$. The first 12 records are grouped into K_1 and the rest are grouped into K_2. Table III shows the comparison between the two cluster structures. The overall F-measure is 0.85 and the loss of cluster quality is 1-0.85=0.15 which is low, suggesting that the generalized data in Table II preserves the cluster structure. ■

D. Beyond Anonymity and F-measure

The framework can flexibly adopt existing clustering, anonymization, and evaluation methods. For example, the generalization can be performed by other anonymization algorithms for classification analysis in [3][6][7][9][12]. The framework can achieve other privacy notions, such as ℓ-diversity [13] and confidence bounding [24][25]. This extension will require modifying the quality function in their algorithms to bias on generalizations that can discriminate class labels. The evaluation method could be entropy [16] or any ad-hoc method defined by the data holder. The data holder could also examine the cluster structures using some visualization tools.

V. Experimental Study

We study the usefulness of the generalized data for cluster analysis. The data set is a publicly available census data set, *Adult*, which was previously used in [3][6][7][9][12][13][22][23][24][25]. There are 45,222 records on 8 categorical and 6 continuous attributes. The continuous attributes are normalized as a standard procedure for many clustering algorithms. Refer to [6][7] for the property and taxonomy of the attributes. For clustering algorithms, we use bisecting and basic k-means [10] due to their popularity. All experiments were carried by an Intel Pentium IV 2.6GHz PC with 1GB RAM.

For the same anonymity threshold, a single QID is always more restrictive than breaking it into multiple QIDs [6][7], so we show the experiment results for single QID only. To ensure that the QID contains attributes that have impact on clustering, we used the C4.5 classifier [14] to rank the attributes. The top ranked attribute is the attribute at the top of the C4.5 decision tree. Then, we remove the top attribute and repeat the process to determine the rank of the remaining attributes. Note, the ranking depends on the raw cluster structure, which depends on the cluster number and clustering algorithm for extracting it. In our experiments, QID contains the top 9 attributes.

A. Homogenous Clustering

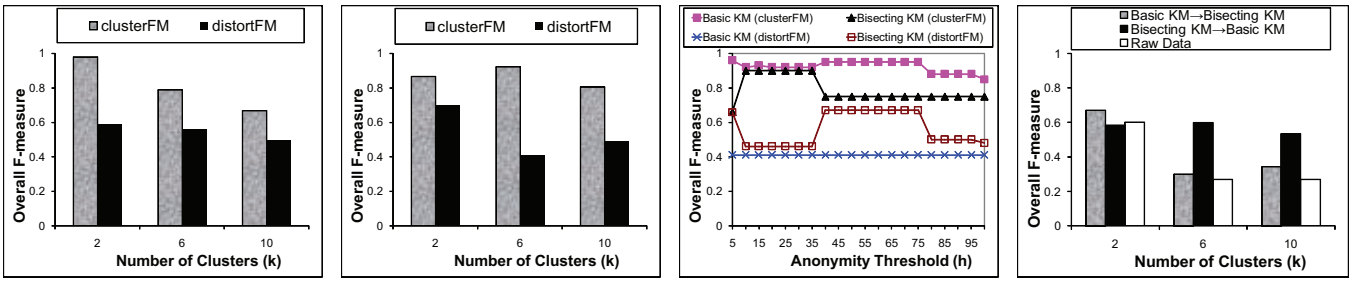
We first present the results for the case where the same clustering algorithm is applied in both Step 1 and Step 3, corresponding to the scenario that the recipient applies the same clustering algorithm as the one used by the data holder. A naive approach to the proposed privacy problem is to ignore the cluster structure and simply anonymize the data using some general purpose anonymization methods [11][15]. We compare our cluster structure-guided anonymization approach with the general purpose anonymization approach by their overall F-measures. To ensure a fair comparison, both approaches employ the same TDS anonymization method [6][7] and the only difference is their quality-guided (score) function. The two overall F-measures are described as follows:

- *clusterFM* denotes the overall F-measure of the cluster structures before and after generalization by *our* cluster structure-guided anonymization approach. Specifically, each specialization maximizes information gain wrt the class labels and minimizes anonymity loss.
- *distortFM* denotes the overall F-measure of the cluster structures before and after generalization by the general purpose anonymization approach that aims at minimizing distortion [15]. Distortion is the number of times a child value is generalized to its parent value in all records.

Figures 3(a) and 3(b) show the averaged *clusterFM* and *distortFM* over anonymity thresholds $5 \leq h \leq 100$ for bisecting and basic k-means where k = 2, 6, 10. $\frac{clusterFM - distortFM}{distortFM}$ represents the *benefit* of our quality-guided approach over the general purpose approach. The benefit, which spans from 24% to 125%, is very significant. F-measure is 1 if two cluster structures are identical, so $\frac{1 - clusterFM}{1} = 1 - clusterFM$ represents the *loss* of cluster quality in order to achieve a given anonymity requirement. The loss spans from 2% to 46%. If the loss is large, e.g., k = 10 in Figure 3(a), the data holder may release an alternative version with a different k. Figure 3(c) displays the overall F-measure over $5 \leq h \leq 100$. Due to limited space, we show the result for $k = 6$ only. *clusterFM* is above 0.7 in 39 out of the 40 test cases, suggesting that there is room for achieving a reasonable level of anonymity ($h \leq 100$) without compromising cluster quality.

B. Heterogenous Clustering

Heterogenous clustering refers to the case that different clustering algorithms are applied in Step 1 and Step 3. It models the scenario that the recipient applies a different clustering algorithm to the generalized data than the one used by the data holder for generalizing the data. We applied bisecting and basic k-means in Step 1 and Step 3 in two different orders, labelled (*Basic KM→Bisecting KM*) and (*Bisecting*



(a) Homogenous: Bisecting k-means (b) Homogenous: Basic k-means (c) Homogenous: $k = 6$ (d) Heterogenous: *clusterFM*

Fig. 3. Overall F-measure for the *Adult* data set

KM→Basic KM), respectively in Figure 3(d). In both cases, compared to the homogenous clustering, there is a drop in the overall F-measure for *clusterFM*. To explain this drop, we measure the overall F-measure of the two cluster structures generated from the raw data without generalization, denoted by *Raw Data* in the figure. Since the overall F-measure of *Raw Data* is also low, the result suggests that the drop is caused by the heterogenous clustering, not by the anonymization.

The above studies suggest that if the recipient applies the same clustering algorithm as the one used for generalizing the data, she will obtain a cluster structure that is more similar to the raw cluster structure because the second clustering could extract the embedded structure preserved in the generalized data. In contrast, if different clustering algorithms are used, the structure preserved by generalization may not be useful to the second clustering due to different search bias.

VI. Conclusions

We studied the problem of releasing person-specific data for cluster analysis while protecting individual privacy. The approach generalizes the unnecessarily specific identifying information but preserves essential cluster structure. Our main contribution is to present a general anonymization framework for properly preserving cluster structures and evaluating the resulting cluster solution. The experimental results on real-life data verify the effectiveness of the approach.

VII. Acknowledgment

The authors would like to thank Mr. Sourav Chakraborty for his contribution in the early stage of this project. Benjamin C. M. Fung's research is supported in part by NSERC Discovery Grants, NSERC Postgraduate Scholarship 50227345, and Concordia University Faculty Start-up Funds. Ke Wang's research is supported in part by NSERC Discovery Grants. Lingyu Wang's research is supported in part by NSERC Discovery Grants N01035.

References

[1] C. C. Aggarwal and P. S. Yu, "A condensation approach to privacy preserving data mining," in *Proc. of the 9th EDBT*, 2004, pp. 183–199.

[2] G. Aggarwal, T. Feder, K. Kenthapadi, R. Motwani, R. Panigrahy, D. Thomas, and A. Zhu, "Achieving anonymity via clustering," in *Proc. of the 25th ACM SIGMOD-SIGACT-SIGART PODS*, June 2006.

[3] R. J. Bayardo and R. Agrawal, "Data privacy through optimal k-anonymization," in *Proc. of the 21st IEEE ICDE*, 2005, pp. 217–228.

[4] CNW Group, "Canadians continue to think personal information not well protected," October 2007, http://www.newswire.ca/en/releases/archive/October2007/17/c5178.html.

[5] B. C. M. Fung, K. Wang, A. W. C. Fu, and J. Pei, "Anonymity for continuous data publishing," in *Proc. of the 11th EDBT*, March 2008.

[6] B. C. M. Fung, K. Wang, and P. S. Yu, "Top-down specialization for information and privacy preservation," in *Proc. of the 21st IEEE ICDE*, April 2005, pp. 205–216.

[7] ——, "Anonymizing classification data for privacy preservation," *IEEE Transactions on Knowledge and Data Engineering (TKDE)*, vol. 19, no. 5, pp. 711–725, May 2007.

[8] A. Inan, S. V. Kaya, Y. Saygin, E. Savas, A. A. Hintoglu, and A. Levi, "Privacy preserving clustering on horizontally partitioned data," *Data & Knowledge Engineering (DKE)*, vol. 63, no. 3, pp. 646–666, 2007.

[9] V. S. Iyengar, "Transforming data to satisfy privacy constraints," in *Proc. of the 8th ACM SIGKDD*, July 2002, pp. 279–288.

[10] L. Kaufman and P. J. Rousseeuw, *Finding Groups in Data: An Introduction to Cluster Analysis*. John Wiley and Sons, March 1990.

[11] K. LeFevre, D. J. DeWitt, and R. Ramakrishnan, "Incognito: Efficient full-domain k-anonymity," in *Proc. of ACM SIGMOD*, 2005, pp. 49–60.

[12] ——, "Workload-aware anonymization," in *Proc. of the 12th ACM SIGKDD*, August 2006.

[13] A. Machanavajjhala, D. Kifer, J. Gehrke, and M. Venkitasubramaniam, "ℓ-diversity: Privacy beyond k-anonymity," *ACM TKDD*, March 2007.

[14] J. R. Quinlan, *C4.5: Programs for Machine Learning*. Morgan Kaufmann, 1993.

[15] P. Samarati, "Protecting respondents' identities in microdata release," *IEEE TKDE*, vol. 13, no. 6, pp. 1010–1027, November-Decemeber 2001.

[16] C. E. Shannon and W. Weaver, *The Mathematical Theory of Communication*. University of Illinois Press, 1963.

[17] L. Sweeney, "k-Anonymity: a model for protecting privacy," in *International Journal on Uncertainty, Fuzziness and Knowledge-based Systems*, vol. 10, no. 5, 2002, pp. 557–570.

[18] U. S. Congress, "Uniting and Strengthening America by Providing Appropriate Tools Required to Intercept and Obstruct Terrorism Act (USA PATRIOT Act)," 2001.

[19] U. S. Department of Homeland Security, "Department of homeland security announces $9 million in information technology grants," September 2004, http://www.dhs.gov/xnews/releases/press_release_0515.shtm.

[20] J. Vaidya and C. Clifton, "Privacy-preserving k-means clustering over vertically partitioned data," in *Proc. of ACM SIGMOD*, 2003.

[21] C. J. van Rijsbergen, *Information Retrieval, 2nd edition*, London, Butterworths, 1979.

[22] K. Wang and B. C. M. Fung, "Anonymizing sequential releases," in *Proc. of the 12th ACM SIGKDD*, August 2006, pp. 414–423.

[23] K. Wang, B. C. M. Fung, and G. Dong, "Integrating private databases for data analysis," in *Proc. of IEEE ISI*, May 2005, pp. 171–182.

[24] K. Wang, B. C. M. Fung, and P. S. Yu, "Template-based privacy preservation in classification problems," in *Proc. of the 5th IEEE ICDM*, Houston, TX, November 2005, pp. 466–473.

[25] ——, "Handicapping attacker's confidence: An alternative to k-anonymization," *Knowledge and Information Systems (KAIS)*, vol. 11, no. 3, pp. 345–368, April 2007.

[26] N. Zhang, "On the communication complexity of privacy-preserving information sharing protocols," in *Proc. of IEEE ISI*, May 2007.