Privacy Preserving Reconstruction-based Techniques and Randomisation-based Methods for Calculating Surveys' Statistics and Participants Sampling in Deliberative Consultations

Piotr Andruszkiewicz

Institute of Computer Science
Warsaw University of Technology
Warsaw, Poland
Email: P.Andruszkiewicz@ii.pw.edu.pl

***Abstract*—In deliberative consultations, the most important are the opinions of residents that want to discuss an important issue. In order to encourage them to participate in such consultations, besides the Internet platform that facilitates the whole process of consultations, privacy preserving techniques should be employed. In this paper, we propose a framework for privacy incorporation in deliberative consultation that will improve eGovernment services provided for digital society. We present the solution for reconstruction-based privacy preserving technique and randomisation-based methods. The proposed framework enables a scientist to prepare a list of candidates and calculate statistics over privacy preserved survey data in deliberative consultations.**

Keywords–Privacy Preserving; Reconstruction-based techniques; Randomisation-based methods; deliberative consultations.

I. INTRODUCTION

In deliberative consultations [1], residents discuss issues important for them. An example could be a deliberative consultation run by a local government in order to find and understand opinions of residents about the desired place of building a new elementary school.

In the era of digital society, organizers provide Internet portals that facilitate the process of gathering opinions of residents. As a starting point, residents provide their characteristics in order to be invited to a deliberative consultation that is of their interest. Moreover, one of the most important methods of gathering data during consultations is providing electronic surveys, especially Internet surveys. Residents may give their opinions through the surveys.

In order to encourage candidates to provide their characteristics, participate in a survey and provide true opinions, privacy should be preserved. To this end, several techniques for incorporating privacy in data mining can be employed. These methods are also helpful in statistical tasks, e.g., mean calculation, that are often used in analysis of data collected in surveys. In [2] we performed the analysis of applicability in deliberative consultations of the following privacy preserving techniques: heuristic-based, reconstruction-based, and cryptography-based [3]. We showed that reconstruction-based privacy preserving technique is useful for deliberative consultations and can provide adequate level of privacy in order to encourage residents to participate in surveys which makes consultations valuable.

In this paper, we propose how to use reconstruction-based techniques and randomisation-based methods for deliberative consultations; namely, for calculating statistics over data colected by surveys and calculating a list of candidates for a deliberative consultation. In our solution we assume that data that was only distorted by means of randomistaion-based methods is collected and stored as a centralised database. The database describes candidates' characteristics and results of surveys.

The remainder of this paper is organized as follows: in Section II, we discuss works related to our task. Section III presents the privacy preserving data mining solutions important in the context of deliberative consultations. In Section IV we propose the solution for consultations with the usage of reconstruction-based technique. Finally, Section V summarises the conclusions of the study and outlines future avenues to explore.

II. RELATED WORK

In this section, we present literature review of privacy preserving classification as it is the field closest to our task and we adopt some of the algorithms presented in literature in order to create a solution for the task in question.

Privacy preserving classification has been extensively discussed in literature [4]–[9].

The pioneer work in privacy preserving classification for centralised data was [10], where R. Agrawal and R. Srikant proposed how to build a decision tree over centralised data distorted with the randomisation-based method (except the target/class attribute) and then classify not distorted data with this decision tree. In this solution, they also presented the algorithm called AS (Agrawal-Srikant) for a probability distribution reconstruction for continuous attributes, which estimates an original probability distribution based on distorted samples (details about the algorithm AS can be found in Section III-B2).

Paper [11] extends the AS (Agrawal Srikant) algorithm and presents the EM (Expectation Maximisation) reconstruction algorithm, which does not take into account nominal attributes either (for details refer to Section III-B3).

Randomised Response technique for related-question model was presented in [12]. It allows creating a decision

 ISBN: 978-1-61208-521-0

tree but only for nominal attributes. Randomised Response technique for unrelated-question model was discussed in [13], [14] and applied in building naïve Bayes classifier.

The solution we proposed in [15] differs from those above, because it enables a miner to classify centralised perturbed data containing simultaneously continuous and nominal attributes by means of randomisation-based methods to preserve privacy on an individual level. This approach uses the EM/AS (Expectation Maximisation/Agrawal Srikant) algorithm (described in details in Section III-B5) to reconstruct a probability distribution for nominal attributes and the ARVeSNA (Algorithm for Assigning Reconstructed Values to Samples for Nominal Attributes) algorithm (please refer to Section III-B7) for assigning reconstructed values to samples for this type of attributes to build a decision tree simultaneously with continuous attributes.

In [16], we proposed the EQ (the abbreviation comes from *system of EQuations*) algorithm (details can be found in Section III-B6) for reconstructing a probability distribution of nominal attributes. The algorithm achieves better results, especially for high level of privacy, i.e., low probability of retaining an original value of a nominal attribute.

Our work is different from the above mentioned proposals as it focuses on calculation of statistics based on privacy preserved centralised database and sampling participants for deliberative consultations. To this end we adopt algorithms developed for privacy preserving classification. We differ from Randomised Response technique for related-question and unrelated-question models because we assume that all surveys' participants answer the same questions.

III. Privacy Preserving Data Mining Preliminaries

A. Randomisation-based Methods

For detailed description of randomisation-based methods used in Privacy Preserving Data Mining please refer to [17].

B. Algorithms for Distribution Reconstruction and for Assigning Reconstructed Values to Samples

The algorithms for distribution reconstruction of both nominal and continuous attributes are described in this section. Moreover, the algorithms for assigning reconstructed values to samples for nominal and continuous attributes are presented. The definition of information loss in reconstruction is introduced, as well.

1) Information Loss: The lack of precision in the reconstruction of a probability distribution is called information loss. It is defined as follows [11]:

Definition Information loss $\mathcal{I}(f_X, \hat{f}_X)$ equals half of the expected value of L_1 norm between the original probability distribution f_X and its estimate $\hat{f}_X$.

$$\mathcal{I}(f_X, \hat{f}_X) = \frac{1}{2} E[\int_{\Omega_X} \mid f_X - \hat{f}_X \mid]$$

Information loss $\mathcal{I}(f_X, \hat{f}_X)$ lies between 0 and 1. $\mathcal{I}(f_X, \hat{f}_X) = 0$ means the perfect reconstruction, and $\mathcal{I}(f_X, \hat{f}_X) = 1$ implies that there is no overlap between the original distribution and its estimate.

2) AS Algorithm for Probability Distribution Reconstruction of Continuous Attributes: The algorithm AS for a probability density function reconstruction for continuous attributes distorted with the randomisation-based method was proposed in [10].

The algorithm solves the following problem:

Original values $x_1, x_2, ..., x_n$ of a one-dimensional distribution are the realisation of n independent random variables $X_1, X_2, ..., X_n$ with the same distribution as the variable X. To hide information, n independent random variables $Y_1, Y_2, ..., Y_n$ with the same distribution as the random variable Y have been used. Given $x_1 + y_1, x_2 + y_2, ..., x_n + y_n$ (y_i is the realisation of the random variable Y_i) and cumulative distribution function F_Y for the variable Y, a cumulative distribution function F_X for the random variable X is to be estimated.

The solution to the given problem is as follows:

Let w_i be the value of $X_i + Y_i$, thus $w_i = x_i + y_i$. The individual values x_i and y_i are not known, only their sums are revealed. Assuming that the probability density function f_X for variable X and f_Y for Y are known, Bayes rule [18] can be used to estimate the posterior (cumulative) distribution function F'_{X_1} for the variable X_1. The posterior distribution function F'_{X_1} can be written as follows:

$$F'_{X_1}(a) = \int_{-\infty}^{a} f_{X_1}(z|X_1 + Y_1 = w_1)dz, \qquad (1)$$

where $F'_{X_1}(a)$ is the estimator of the posterior (cumulative) distribution function $F_{X_1}(a)$.

Using Bayes rule:

$$F'_{X_1}(a) = \int_{-\infty}^{a} \frac{f_{X_1+Y_1}(w_1|X_1 = z) f_{X_1}(z)}{f_{X_1+Y_1}(w_1)} dz. \qquad (2)$$

After additional calculations [10] the posterior density function f'_X is obtained by differentiating F'_X:

$$f'_X(a) = \frac{1}{n} \sum_{i=1}^{n} \frac{f_Y(w_i - a) f_X(a)}{\int_{-\infty}^{\infty} f_Y(w_i - z) f_X(z) dz}. \qquad (3)$$

Having a large number of samples, f'_X should correspond to the original probability density function f_X.

To estimate f'_X, the knowledge of f_Y and f_X is needed. f_Y is known, because the distorting distribution function is known for a miner. As the original probability density function f_X is unknown, a uniform distribution is assumed as an initial estimate of density function and then refined in an iterative way by applying (3). See Figure 1 for details.

Details about the calculation complexity reduction can be found in [10].

To stop an iterate reconstruction, three possible stopping criteria were proposed in [10].

The first criterion is met when the reconstructed distribution is statistically the same as the original distribution. To check the similarity of distributions, for instance, χ^2 measure (details about χ^2 can be found in [19]) can be used. This

$f_X^0 :=$ uniform distribution
$j := 0$ // iteration number
repeat
$f_X^{j+1}(a) = \frac{1}{n} \sum_{i=1}^{n} \frac{f_Y(w_i - a) f_X^j(a)}{\int_{-\infty}^{\infty} f_Y(w_i - z) f_X^j(z) dz}$
$j := j + 1$
until(stopping criterion met)

Figure 1. The AS algorithm.

$Pr(X = v_p)^0 := \frac{1}{k}, p = 1, ..., k$
$j := 0$ //iteration number
repeat
$Pr(X = v_p)^{j+1} = \frac{1}{n} \sum_{s=1}^{n} \frac{Pr(v_p \to X(s)) Pr^j(X = v_p)}{\sum_{t=1}^{k} Pr(v_t \to X(s)) Pr^j(X = v_t)}$
$j := j + 1$
until(stopping criterion met)

Figure 2. The EM/AS nominal attribute probability distribution reconstruction algorithm.

criterion could be used only for testing, because the original distribution is not known in practice.

The second solution is to compare the randomised current estimate of the original distribution with the distorted distribution used for the reconstruction and stop when these two distributions are statistically the same. This criterion assumes that the current estimate which is close enough to the original distribution should be the same after the distortion as the distorted distribution used for the reconstruction. As stated in [10], the difference between two distorted distributions is not a reliable indicator.

The last approach is to compare two consecutive estimates of the original distribution. When the difference is small enough, the process is completed. 1% of the threshold of χ^2 test was used in [10].

As stated in [11], the AS algorithm may not always converge and even it converges, there is no guarantee that it gives a reasonable estimate of the original distribution. There was no proof given for that statement and this issue was not mentioned in [10].

3) Algorithm EM for Probability Distribution Reconstruction of Continuous Attributes: The algorithm for a probability density function reconstruction for continuous attributes distorted by means of the the randomisation-based method was proposed in [11], as well. The algorithm was called EM by the authors. The problem to be solved is the same as for the AS algorithm.

The details about the EM algorithm and the proof that it converges can be found in [11]. The authors of the EM algorithm stated that it is theoretically the best algorithm and having a large set of distorted samples, the EM algorithm can reconstruct the original distribution with little or without information loss [11]. The definition of information loss can be found in Section III-B1.

4) Assigning Reconstructed Values to Samples for Continuous Attributes: The algorithm for assigning reconstructed values to samples for continuous attributes was presented in [10]. We describe this algorithm in this section.

Let $I_1, ..., I_m$ denote m intervals and $N(I_k)$ be the number of samples in I_k interval. Samples should be sorted in an ascending order and assigned to consecutive intervals as follows: $N(I_1)$ first samples are assigned to the first interval I_1, the next $N(I_2)$ samples to the second interval I_2, etc.

5) EM/AS Algorithm for Probability Distribution Reconstruction of Nominal Attributes: In [15], we proposed the EM/AS algorithm for reconstructing a probability distribution of a nominal attribute.

The EM/AS algorithm is based on two algorithms: AS proposed in [10] and its extension EM presented in [11]. Both algorithms reconstruct a probability distribution of continuous attributes.

To reconstruct probability distribution of a nominal attribute, both EM and AS algorithms were modified to obtain the EM/AS (Figure 2). The modifications of both algorithms (AS and EM) give the same result.

The algorithm solves the following problem: a nominal attribute X has the possible values $v_1, v_2, v_3, ..., v_k$ and n samples. Value for each sample is modified according to a probability $Pr(v_p \to v_r)$ (a probability that a value v_p will be changed to a value v_r). $X(s)$ means a value of an attribute X for a sample s. An original probability distribution of an attribute X should be reconstructed.

The algorithm starts with the uniform distribution and calculates the estimate of the probability distribution in every iteration.

Stopping criterion is the same as for the AS and EM algorithms (the algorithm is stopped when the difference between successive estimates of the original probability distribution becomes small, as little as 1% of the threshold of the χ^2 test).

6) EQ Algorithm for Probability Distribution Reconstruction of Nominal Attributes: In [16], we proposed the EQ algorithm, the name of the algorithm comes from the phrase *system of EQuations*, that reconstructs the probability distribution of nominal attributes and can be used instead of the EM/AS algorithm. The EQ algorithm outperforms the EM/AS, especially for high levels of privacy [16].

The problem to be solved is the same as for the EM/AS algorithm: there are a nominal attribute X with the possible values $v_1, v_2, v_3, ..., v_k$ and n samples. A value for each sample is modified according to a probability $Pr(v_p \to v_r)$ (a probability that a value v_p will be changed to a value v_r) and we want to reconstruct an original probability distribution of an attribute X.

Let us assume that there is an attribute *Colour* with 3 values: $v_1 = green$, $v_2 = blue$, and $v_3 = black$.

For the original value of the attribute, e.g., $green$, the probability $Pr(v_1 \to v_1)$ that the value will be the same after the modification is known, as well as the probability of changing the value from $green$ to $blue$ and from $green$ to $black$. Moreover, when the value of the attribute after the distortion is, e.g., $green$, the original value was one of the three possible values: $green$, $blue$, and $black$ and all the probabilities $Pr(v_1 \to v_1)$, $Pr(v_2 \to v_1)$, $Pr(v_3 \to v_1)$ how the value has become $green$ are known.

Let Z be the attribute after the modification with the possible values $v_1, v_2, v_3, ..., v_k$. In the example, the attribute

 ISBN: 978-1-61208-521-0

Z has 3 values: $green$, $blue$, and $black$ and the following equation can be written:

$$P(Z = green) = a_{1,1}P(X = green) + a_{1,2}P(X = blue) + a_{1,3}P(X = black),$$

where $a_{s,p} = Pr(v_p \rightarrow v_s)$. For colours $blue$ and $black$ the similar equations can be written:

$$P(Z = blue) = a_{2,1}P(X = green) + a_{2,2}P(X = blue) + a_{2,3}P(X = black)$$
$$P(Z = black) = a_{3,1}P(X = green) + a_{3,2}P(X = blue) + a_{3,3}P(X = black).$$

Now there are 3 equations and 3 unknown variables ($P(X = green)$, $P(X = blue)$, $P(X = black)$), thus the system of linear equations can be solved.

In general there is the following system of k equations:

$$P(Z = v_1) = a_{1,1}P(X = v_1) + a_{1,2}P(X = v_2) + \cdots + a_{1,k}P(X = v_k)$$
$$P(Z = v_2) = a_{2,1}P(X = v_1) + a_{2,2}P(X = v_2) + \cdots + a_{2,k}P(X = v_k)$$
$$\vdots$$
$$P(Z = v_k) = a_{k,1}P(X = v_1) + a_{k,2}P(X = v_2) + \cdots + a_{k,k}P(X = v_k)$$

with k unknown variables.

Let $\mathbf{X}$ be the column vector with elements $x_1, ..., x_k$, where $x_i = P(X = v_i)$ and $\mathbf{Z}$ be the column vector with elements $z_1, ..., z_k$, where $z_i = P(Z = v_i)$. Let $\mathbf{P}$ be the matrix of retaining/changing values of a nominal attribute. We can rewrite the system of equations in the matrix form as:

$$\mathbf{Z} = \mathbf{PX} \tag{4}$$

To find values of $P(X = v_i),\ i = 1, \ldots, k$, we need to solve (4). We can solve it by left multiplying both sides by inverted $\mathbf{P}$, i.e., $\mathbf{P}^{-1}$ (only if inverted $\mathbf{P}$ exists).

Nonexistence of the inverted matrix is not troublesome because the number of values of a nominal attribute is known before collecting data starts and a non-singular matrix $\mathbf{P}$ can be chosen to guarantee the existence of inverted $\mathbf{P}$ matrix.

7) ARVeSNA Algorithm for Assigning Reconstructed Values to Samples for Nominal Attributes: We proposed the algorithm for assigning reconstructed values to samples for nominal attributes in [20] and describe this algorithm in this section.

Having reconstructed a probability distribution of a nominal attribute, reconstructed values can be assigned to samples.

The algorithm solves the following problem:

Since modified values of a nominal attribute are given, the probability distribution of a modified attribute (i.e., $P(Z = v_i),\ i = 1, \ldots, k$) and the number of all samples n are known. The reconstructed probability distribution ($P(X = v_i),\ i = 1, \ldots, k$) is estimated. The aim is to assign reconstructed values to samples taking into account the reconstructed probability distribution.

In order to solve this problem, the number of distorted samples ($n_Z(v_i)$) is counted separately for each value of an attribute and the number of original samples ($n_X(v_i) = P(X = v_i)n$) is estimated.

TABLE I. THE EXAMPLE OF THE ORIGINAL DATABASE.

Id	Salary	Age	Sex	Credits status	Children
1	1000	35	M	none	N
2	1500	37	F	overdue	Y
3	5000	41	M	present	N
4	3000	44	M	repaid	N
5	4200	50	F	repaid	N
6	2000	28	F	none	N
7	1000	30	M	none	Y

TABLE II. THE EXAMPLE OF THE DISTORTED DATABASE WITH UNIFORM DISTORTION DISTRIBUTION $\langle -500, 500\rangle$ FOR SALARY, $\langle -10, 10\rangle$ FOR AGE AND $p = 0.6$ FOR SEX AND CREDITS STATUS ATTRIBUTES.

Id	Salary	Age	Sex	Credits status	Children
1	1353.32	33.42	M	repaid	N
2	1611.83	40.64	M	overdue	Y
3	5428.27	51.27	M	present	N
4	2573.22	39.51	F	none	N
5	4145.89	42.67	M	repaid	N
6	2258.34	38.72	F	none	N
7	1054.03	36.65	M	overdue	Y

Then the difference, called $\delta(v_i)$, between $n_Z(v_i)$ and $n_X(v_i)$ is calculated. $\delta(v_i) > 0$ means that there are too many samples because there are more samples with distorted value of v_i than the reconstructed number of samples for the value v_i. A sample corresponding to a positive value of $\delta(v_i)$ is found and assigned with a reconstructed value v_j for which a value of $\delta(v_j)$ is negative and the reconstructed value v_j has the highest probability to be distorted to the value v_i. Values of corresponding $\delta(v_i)$ and $\delta(v_j)$ are updated and the process is continued until all values of $\delta(v_i),\ i = 1, \ldots, k$ are zero.

Having completed the process, samples with the reconstructed values are assigned according to an original (reconstructed) probability distribution.

IV. CALCULATING SURVEYS' STATISTICS AND PARTICIPANTS SAMPLING

We define two tasks in deliberative consultations that involve calculations on data with preserved privacy by means of reconstruction-based techniques and randomisation-based methods; namely, calculating surveys' statistics and participants sampling.

A. Calculating Surveys' Statistics

In the task of calculating statistics from survey's data, we assume that there is a centralised database that is collected by means of electronic surveys. A participant provides an answer to each question in the survey. The answer is an attribute value. Thus, we can state that the database consists of: a definition of attributes and its values.

Each attribute describes possible answer's values for a survey's question. For example, an attribute that describes a question 'Do you have a car?' is a binary attribute with possible values: *yes* or *no*. The possible types of attributes are: binary, nominal, ordinal, and continuous.

Values of attributes are distorted answers provided by participants. Thus, if we have n participants and k questions,

we have n values for each of k attributes. According to randomisation-based methods values of attributes are distorted at a client side; that is, participant's side, with one of the possible methods. For binary attributes we may use basic randomisation factor method or distortion with a matrix of retaining/changing values of a nominal attribute (for more details please refer to [2]). For ordinal attributes we may use a modified matrix of retaining/changing values of an attribute described in [21]. For continuous attributes the additive perturbation method [10], multiplicative perturbation [22], and the retention replacement perturbation [23] may be employed.

Tables I and II show the example databases that could be results of electronic survey after a deliberative consultation. Table I presents the original, not distorted, database that could be a real result of a survey if there is no privacy preserving methods applied. The example database that could be a result of a survey with privacy preserved by means of the randomisation-based method is shown in Table II. If we use the randomisation-based method in real applications, the original database (Table I) does not exist, only distorted values (Table II) are stored. *Id* attribute is not necessary and is shown to ease the process of comparing both databases.

In the aforementioned task, there is a table with distorted values, like in Table II and a scientist wants to calculate some statistics about participants of a survey. The statistics could be the number of participants that meet a specific condition which is based on gathered data, e.g., the number of participants that have children and are at least 30 years old. Another statistics to calculate is mean of some attribute for all participants or participants that meet a specific condition. For instance, mean salary for participants that have children and are at least 30 years old. Last but not least a scientist may want to see a distribution of an attribute for all participants or a group of participants.

1) Calculating Number of Participants that Meet Specific Condition: Fist we define a condition that a scientist may create in order to choose a group of participants. Let C be a condition that participants should meet. Let us assume that $c_1, c_2, ..., c_m$ are subconditions and form condition C, i.e., $C = c_1 \wedge c_2 \wedge ... \wedge c_m$. c_i condition is a condition that is based on one attribute. The possible types of this conditions are: $v_{a_j} > t$, $v_{a_j} \leq t, t_1 \leq v_{a_j} < t_2$ for continuous attributes, $v_{a_j} \in \{v_1, v_2, ..., v_l\}$ for binary, ordinal, and nominal attributes, where v_{a_j} is a value of attribute a_j, t is a known threshold, $v_1, v_2, ..., v_l$ are possible values of an attribute.

Let us consider that condition $C = c_1$ and c_1 is of a form $t_1 \leq v_{a_j} < t_2$ and is based on a continuous attribute. In this case we can choose intervals in a way that their end/begin in points t_1 and t_2 and apply AS or EM algorithm. Let us assume that we choose the following intervals: i_1 that starts in $-\infty$ and ends in t_1, i.e., $i_1 = (-\infty; t_1)$, $i_2 =< t_1; t_2)$, $i_3 =< t_2; \infty)$. The number of intervals need not to be limited to 3, however, it is important that they should be intervals that end/begin in t_1 and t_2.

The AS or EM algorithms estimate a distribution of values of attribute over intervals. The number of participants can be obtained by multiplying a probability for an interval by the number of participants. Thus the number of values that lie in each interval N_1, N_2, N_3 is known. If we assume that $t_1 < t_2$, then N_2 is the number of participants that meet condition C. Otherwise, $N_1 + N_3$ is the right number. For conditions of form $v_{a_j} > t$, $v_{a_j} \leq t$ the solution is analogous.

If c_1 is based on a nominal attribute with k possible values and is of form $v_{a_j} \in \{v_1, v_2, ..., v_l\}$ EM/AS or EQ algorithm can be employed. The output of the algorithm is the number of participants for each possible value of an attribute. As shown in the following equation, for condition $v_{a_j} \in \{v_1, v_2, ..., v_l\}$ we need to sum all numbers for values that are present in the condition; that is $v_{a_j} \in \{v_1, v_2, ..., v_l\}$.

$$N_{\{v_1, v_2, ..., v_l\}} = \sum_{v_i \in \{v_1, v_2, ..., v_l\}} N_{v_i} \quad (5)$$

where:

- $N_{\{v_1, v_2, ..., v_l\}}$ is the estimated number of samples for which the original attribute has one of values in this set $\{v_1, v_2, ..., v_l\}$,
- N_{v_1} is the estimated number of samples for which the original attribute has a value v_1.

In order to calculate the number of participants that meet condition C with more than one subcondition, $C = c_1 \wedge c_2 \wedge ... \wedge c_m$, we need to calculate the number of participants in an iterative manner (see Figure 3). For the first subcondition we calculate the number of participants that meet subcondition c_1 as shown in this section. Then, we choose participants that meet this subcondition. For binary, nominal and ordinal attributes we can employ ARVeSNA algorithm (Section III-B7, [20]). For continuous attributes we can apply the algorithm described in Section III-B4 that sorts participants in the ascending order over a condition attribute. Then the algorithm assigns the estimated number of sorted participants to each interval. As a result, we obtain a set P_{c_1} of participants that meet a subcondition c_1. In the next iteration we start with P_{c_1} set of participants and apply a subcondition c_2. The result of this iteration is a set $P_{c_1 \wedge c_2}$ of participants that meet condition $c_1 \wedge c_2$. Then we proceed to the next iteration until $P_{c_1 \wedge c_2 \wedge ... \wedge c_m}$ is obtained and hence the number of participants that meet condition C. In the last iteration ARVeSNA or the algorithm for continuous attributes that chooses a subset of participants according to calculated distribution need not to be aplied because we need only the number of participants that meet condition C and the list of participants is not necessary.

As an example, we will analyse the calculation of the number of participants who meet the following conditions: females at most 30 years old.

Let us assume that the Table II contains the distorted results of a survey. We will use the algorithm presented in Figure 3. The first condition is *Sex* = F. In order to find the estimated number of objects that meet this condition we use EM/AS or EQ algorithm. Let us assume that the result of EM/AS or EQ algorithm is 3. Based on Table II without taking into account the distortion we would obtain the number 2. Then, by the means of ARVeSNA algorithm we assign participants to values of the attribute *Sex* in this case. Let us assume that the participants 2, 5 and 6 are assigned. The second condition is *Age* $\leq$ 30. The considered attribute is continuous thus we use EM or AS algorithm. It gives the number of participants at most 30 years old within participants 2, 5 and 6. The result is 1. If we used data from Table II directly without taking into

INPUT: m // number of subconditions
INPUT: $C = c_1 \wedge c_2 \wedge ... \wedge c_m$ // condition to be met
INPUT: P // set of participants
OUTPUT: $P_{c_1 \wedge c_2 \wedge ... \wedge c_m}$ // set of participants that meet $c_1 \wedge c_2 \wedge ... \wedge c_m$ condition
OUTPUT: $N_{c_1 \wedge c_2 \wedge ... \wedge c_m}$ // number of participants that meet $c_1 \wedge c_2 \wedge ... \wedge c_m$ condition

for i = 1 **to** m **do**
  if c_i is based on continuous attribute **then**
    prepare intervals:
    $i_1 = (-\infty; t_1)$, $i_2 =< t_1; t_2)$, $i_3 =< t_2; \infty)$
    calculate distribution with AS or EM algorithm
    assign participants to intervals
    choose $P_{c_1 \wedge ... \wedge c_i}$
    calculate $N_{c_1 \wedge ... \wedge c_i}$
  elseif // binary, nominal, ordinal attributes
    calculate distribution with EM/AS or EQ algorithm
    assign participants to attribute values (ARVeSNA)
    choose $P_{c_1 \wedge ... \wedge c_i}$
    calculate $N_{c_1 \wedge ... \wedge c_i}$
  end
end

Figure 3. The list and the number of participants that meet a specific condition calculation algorithm.

account that values are distorted, we would conclude that there were no females at most 30 years old.

2) Mean Calculation: Considering the mean of continuous attributes and the additive perturbation the calculations are the same as for not distorted data if the distorting distribution with mean equal to 0 is used. The type of a distribution, e.g., uniform, normal, makes no difference. A distribution with mean equal to 0 does not statistically change the mean of attribute's values. The mean can be calculated for an arbitrary set of values of an attribute. Therefore, a scientist is able to calculate a mean if a group of participants is chosen. To this end, a scientist may use the algorithm presented in Figure 3 in order to find a set of participants that meet a specific condition and then calculate the mean.

B. Participants Sampling

In participants sampling, a scientist chooses a set of participants that take part in a deliberative consultation regarding participants' characteristics. Let us assume that all candidates provide information about them and the randomised-based method is applied, hence, only distorted data is stored. Table II may represent such characteristics provided by candidates. A scientist wants to find a condition that chooses a right group of people to be involved in a deliberative consultation. Hence, algorithm presented in Figure 3 can be used to perform this task, since it provides a number and a list of candidates that meet a specific condition.

If we assume that Table II represents characteristics of participants and we want to find the list of candidates for consultations that are females at most 30 years old, we may use the example from Section IV-A1 to illustrate participants samplig. However, in this case we need to perform the last step of ARVeSNA algorithm; that is, assign values of attribute *Age* to participants.

V. Conclusion and Future Work

In this work, we proposed a framework for reconstruction-based techniques and randomisation-based methods application in deliberative consultations. The solution for calculating statistics over privacy preserved survey data and candidates' characteristics has been presented. The proposed framework lets a scientist apply privacy preserving in real deliberative consultations.

In future work, we plan to investigate the possibility of k-anonymity application in a hybrid solution that combines aggregation, reconstruction-based technique and k-anonymity approach.

The incorporation of the presented framework in the system for deliberative consultations that is being under development is planned also.

Acknowledgment

This research has been supported by the National Centre for Research and Development under grant No SP/I/1/77065/10 and the Institute of Computer Science, Warsaw University of Technology under Grant No. II/2015/DS/1.

References

[1] J. Abelson et al., "Deliberations about deliberative methods: issues in the design and evaluation of public participation processes," Social science & medicine, vol. 57, no. 2, 2003, pp. 239–251.

[2] P. Andruszkiewicz, "Privacy preserving data mining for deliberative consultations," in (accepted for) Hybrid Artificial Intelligent Systems - 11th International Conference, HAIS 2016, Seville, Spain, April 18-20, 2016.

[3] V. S. Verykios et al., "State-of-the-art in privacy preserving data mining." SIGMOD Record, vol. 33, no. 1, 2004, pp. 50–57.

[4] L. Xiong, S. Chitti, and L. Liu, "Mining multiple private databases using a knn classifier," in SAC '07: Proceedings of the 2007 ACM symposium on Applied computing, 2007, pp. 435–440.

[5] M. R. Keyvanpour and S. S. Moradi, "A perturbation method based on singular value decomposition and feature selection for privacy preserving data mining," IJDWM, vol. 10, no. 1, 2014, pp. 55–76. [Online]. Available: http://dx.doi.org/10.4018/ijdwm.2014010104 [accessed: 2016-03-18]

[6] P. Andruszkiewicz, "Hierarchical combining of classifiers in privacy preserving data mining," in Hybrid Artificial Intelligence Systems - 9th International Conference, HAIS 2014, Salamanca, Spain, June 11-13, 2014. Proceedings, ser. Lecture Notes in Computer Science, M. M. Polycarpou, A. C. P. L. F. de Carvalho, J. Pan, M. Wozniak, H. Quintián, and E. Corchado, Eds., vol. 8480. Springer, 2014, pp. 573–584. [Online]. Available: http://dx.doi.org/10.1007/978-3-319-07617-1 [accessed: 2016-03-18]

[7] X. Li, Z. Yan, and P. Zhang, "A review on privacy-preserving data mining," in 14th IEEE International Conference on Computer and Information Technology, CIT 2014, Xi'an, China, September 11-13, 2014. IEEE, 2014, pp. 769–774. [Online]. Available: http://dx.doi.org/10.1109/CIT.2014.135 [accessed: 2016-03-18]

[8] P. Andruszkiewicz, "Frequent sets discovery in privacy preserving quantitative association rules mining," in Hybrid Artificial Intelligent Systems - 10th International Conference, HAIS 2015, Bilbao, Spain, June 22-24, 2015, Proceedings, ser. Lecture Notes in Computer Science, E. Onieva, I. Santos, E. Osaba, H. Quintián, and E. Corchado, Eds., vol. 9121. Springer, 2015, pp. 3–15. [Online]. Available: http://dx.doi.org/10.1007/978-3-319-19644-2 [accessed: 2016-03-18]

[9] J. Hamm, A. C. Champion, G. Chen, M. Belkin, and D. Xuan, "Crowd-ML: A privacy-preserving learning framework for a crowd of smart devices," CoRR, vol. abs/1501.02484, 2015. [Online]. Available: http://arxiv.org/abs/1501.02484 [accessed: 2016-03-18]

[10] R. Agrawal and R. Srikant, "Privacy-preserving data mining," in SIGMOD Conference, W. Chen, J. F. Naughton, and P. A. Bernstein, Eds. ACM, 2000, pp. 439–450.

 ISBN: 978-1-61208-521-0

[11] D. Agrawal and C. C. Aggarwal, “On the design and quantification of privacy preserving data mining algorithms,” in PODS ’01: Proceedings of the twentieth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems, 2001, pp. 247–255.

[12] W. Du and Z. Zhan, “Using randomized response techniques for privacy-preserving data mining.” in KDD, 2003, pp. 505–510.

[13] J. Z. Zhan and S. Matwin, “Privacy-preserving data mining in electronic surveys,” in ICEB, J. Chen, Ed. Academic Publishers/World Publishing Corporation, 2004, pp. 1179–1185.

[14] ——, “Privacy-preserving data mining in electronic surveys,” I. J. Network Security, vol. 4, no. 3, 2007, pp. 318–327.

[15] P. Andruszkiewicz, “Privacy preserving classification for continuous and nominal attributes,” in Proceedings of the 16th International Conference on Intelligent Information Systems, 2008.

[16] ——, “Probability distribution reconstruction for nominal attributes in privacy preserving classification,” in ICHIT ’08: Proceedings of the 2008 International Conference on Convergence and Hybrid Information Technology. Washington, DC, USA: IEEE Computer Society, 2008, pp. 494–500.

[17] C. C. Aggarwal, “Privacy-preserving data mining,” in Data Mining. Springer International Publishing, 2015, pp. 663–693.

[18] M. Fisz, Probability Theory and Mathematical Statistics. New York: John Wiley and Sons, 1963.

[19] L. M. Surhone, M. T. Timpledon, and S. F. Marseken, Pearson’s Chi-Square Test. Beau Bassin: Betascript Publishers, 2010.

[20] P. Andruszkiewicz, “Privacy preserving data mining on the example of classification (in Polish),” Master’s thesis, Warsaw University of Technology, 2005.

[21] ——, “Privacy preserving classification for ordered attributes,” in Man-Machine Interactions, ser. Advances in Soft Computing, J. F. P. U. S. Krzysztof A. Cyran, Stanisaw Kozielski and A. Wakulicz-Deja, Eds., vol. 59/2009. Springer, 2009, pp. 353–360.

[22] J. J. Kim and W. E. Winkler, “Multiplicative noise for masking continuous data,” Statistical Research Division, US Bureau of the Census, Washington D.C., Tech. Rep., 2003.

[23] R. Agrawal, R. Srikant, and D. Thomas, “Privacy preserving olap,” in SIGMOD ’05: Proceedings of the 2005 ACM SIGMOD international conference on Management of data. New York, NY, USA: ACM, 2005, pp. 251–262.

 ISBN: 978-1-61208-521-0