Microfiles as a Potential Source of Confidential Information Leakage

Oleg Chertov and Dan Tavrov

Abstract Cyber warfares, as well as conventional ones, do not only comprise direct military conflicts involving weapons like DDoS attacks. Throughout their history, intelligence and counterintelligence played a major role as well. Information sources for intelligence can be closed (obtained during espionage) or open. In this chapter, we show that such open information sources as microfiles can be considered a potentially important additional source of information during cyber warfare. We illustrate by using real data based example that ignoring issues concerning providing group anonymity can lead to leakage of confidential information. We show that it is possible to define fuzzy groups of respondents and obtain their distribution using appropriate fuzzy inference system. We conclude the chapter with discussing methods for protecting distributions of crisp as well as fuzzy groups of respondents, and illustrate them by solving the task of providing group anonymity of a fuzzy group of "respondents who can be considered military enlisted members with the high level of confidence."

1 Introduction

With the development of appropriate information technologies, the role of open information sources as a way of obtaining confidential information becomes more and more significant. Such technologies include means of processing very large amounts of data, text and data mining methods, hardware and software based ways of obtaining and analyzing information from different sources, to name just a few.

According to the research conducted by the International Data Corporation [1], about 30 % of digital information in the world need protection, and this number will rise to roughly 40 % by 2020.

O. Chertov (✉) · D. Tavrov
National Technical University of Ukraine, Kyiv Polytechnic Institute,
37 Peremohy Prospekt, 03056 Kyiv, Ukraine
e-mail: chertov@i.ua

D. Tavrov
e-mail: dan.tavrov@i.ua

R.R. Yager et al. (eds.), *Intelligent Methods for Cyber Warfare*,
Studies in Computational Intelligence 563, DOI 10.1007/978-3-319-08624-8_4

A *microfile* is a collection of primary data with information about a sample respondent set. Microfiles are constructed using census or other statistical and sociological surveys data, marketing research data, social networks analysis data etc. With the help of the primary microfile data, as opposed to aggregated ones, one can try to obtain answers to questions not foreseen by the microfile creators.

Microfiles can be considered a potentially important source of information during cyber warfare. With their help, it is possible to violate individual or group anonymity. *Anonymity* of an object means that this object is unidentifiable among the set of certain objects [2]. *Individual* data anonymity is a property of information on a single respondent to be unidentifiable within the data set [3, p. 1]. *Group* data anonymity is a condition, under which [4, p. 11] data features that cannot be distinguished while considering individual records only are protected.

Individual anonymity can be violated even when attributes that uniquely identify microfile respondents are removed. For instance, as L. Sweeney showed experimentally in 2001, 97 % of the voters in the state of Massachusetts possess unique combination of birth date (day, month, and year) and nine-digit ZIP code [5]. Appropriate methods for providing individual anonymity were introduced, such as randomization [6], microaggregation [7], data swapping [8], data matrix factorization [9] and singular value decomposition [10], wavelet transforms (WT) [11], etc.

Group anonymity can be violated by analyzing distributions of the microfile data over certain attribute values. For example, Fig. 1 presents the regional distribution of power engineering specialists obtained from the microfile containing results of the 1999 population census in France [12]. The higher the cylinder, the more specialists live in a particular region. Since the French energy sector primarily consists of nuclear stations (78 % of all energy produced in 2011 [13]), the highest number of power engineering specialists occurred exactly in those regions where nuclear power plants are situated (black cylinders in Fig. 1). Therefore, to conceal the site of any secret nuclear research center, one should distort the real regional distribution of French power engineering specialists.

In the literature, several classes of the task of providing group anonymity (TPGA) are distinguished. The quantity TPGA defined as the task of providing anonymity of a respondent group quantity distribution over the set of values (e.g. military personnel regional distribution) was introduced in [14]. In terms of quantity task, it is impossible to solve the task of concealing concentration distribution of respondents. Such tasks are called concentration group anonymity tasks [15]. One of them is the concentration difference task [16], which implies concealing the distribution of the difference between two concentration distributions. The problems of providing group anonymity are most elaborately covered in [4]. The general methodology of providing group anonymity is presented in [3].

Most of existing methods of solving the TPGA deal with the so called *crisp* groups of respondents, i.e. those ones, to which a particular respondent either belongs or not. The membership in such a group can be determined by analyzing values of one or several specific attributes, e.g., “Occupation,” as in the case of French power engineering specialists. To protect anonymity, one can use existing methods, or, in the crudest case, remove appropriate attributes from the microfile.

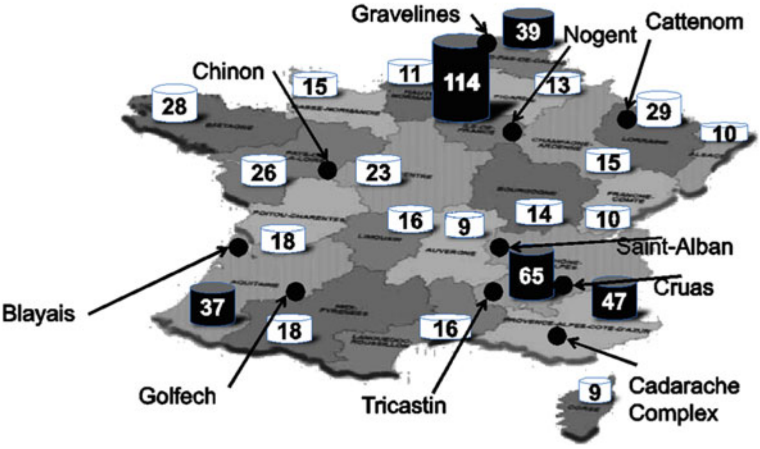

Fig. 1 Regional power engineering specialists distributed according to the microfile containing results of the 1999 French population census

In some cases, however, it is possible to violate group anonymity for a *fuzzy* group of respondents, i.e. the one, to which a respondent can belong only to a certain degree. Whether a respondent belongs to a group, is determined by analyzing values not of some special attributes, but of one or several rather general ones, such as "Age," "Sex," etc. In this chapter, for instance, we discuss a real data based example, in which the fuzzy group consists of people who can be considered military enlisted members with the high level of confidence. The membership in such a group can be deduced from analyzing values of such general purpose attributes as "Age," "Sex," "Black or African American," "Marital Status," "Educational Attainment," and "Hours per Week in 1999." We show that even if group anonymity is provided for a crisp group of military personnel, it might still be possible to retrieve sensitive information from the microfile using the concept of a fuzzy respondent group.

Importance of easily accessed data for retrieving hidden information should not be underestimated. E.g., the famous Russian chemist D. Mendeleyev was able to find out the secret composition of the French powder [17, pp. 353–354] by analyzing annual shipment report of the railroad company that supplied the factory.

Since it is obviously not an option to remove important attributes like "Age," "Sex," etc. from the microfile, appropriate anonymity-providing methods should be developed.

2 General Approach to Violating Group Anonymity

2.1 Group Anonymity Basics

Microdata are the data about respondents (people, households, enterprises etc.). Let **M** denote a (depersonalized) *microfile* with microdata collected in a file of attributive records, which can be viewed as a matrix with rows u_i, $i = \overline{1, \mu}$, corresponding to respondents, and columns w_j, $j = \overline{1, \eta}$, corresponding to attributes.

Let $\mathbf{w}_j$ denote the set of all attribute w_j values. The *vital* set is a subset $\mathbf{V} = \left\{V_1, V_2, \ldots, V_{l_v}\right\}$ of the Cartesian product of *vital* attributes. The elements of $\mathbf{V}$ are the *vital value combinations*. $\mathbf{V}$ enables us to define the respondent group.

We can define linguistic variables [18] L_i corresponding to each microfile attribute. Universes of discourse for L_i consist of the ith microfile attribute values. Values of L_i belong to its term-set $T(L_i)$. The *generalized vital set* $\tilde{\mathbf{V}} = \left\{\tilde{V}_1, \tilde{V}_2, \ldots, \tilde{V}_{l_{\tilde{v}}}\right\}$ is a subset of the Cartesian product of term-sets of all the linguistic variables corresponding to the vital microfile attributes. The elements of $\tilde{\mathbf{V}}$ are the *generalized vital value combinations*.

Let the *parameter set* $\mathbf{P} = \left\{P_1, P_2, \ldots, P_{l_p}\right\}$ be a subset of values corresponding to the *parameter* microfile attribute, which is not vital. The elements of $\mathbf{P}$ are *parameter values*. They enable us to split the microfile $\mathbf{M}$ into *parameter submicrofiles* $\mathbf{M}_1, \ldots, \mathbf{M}_{l_p}$ with μ_j, $j = \overline{1, l_p}$, records in them.

We will denote by $G(\mathbf{V}, \mathbf{P})$ the *group*, i.e. the set consisting of $\mathbf{V}$ and $\mathbf{P}$. We will denote by $\tilde{G}\left(\tilde{\mathbf{V}}, \mathbf{P}\right)$ the *fuzzy group*, i.e. the set consisting of $\tilde{\mathbf{V}}$ and $\mathbf{P}$.

We can determine the membership grade $\mu_{\tilde{G}}(u_i)$ of every respondent u_i, $i = \overline{1, \mu}$, in $\tilde{G}$. We denote the set of all grades by $\tilde{\mathbf{M}}_{\tilde{G}} = \left\{\mu_{\tilde{G}1}, \mu_{\tilde{G}2}, \ldots, \mu_{\tilde{G}q}\right\}$.

By *goal representation* $\Omega\left(\mathbf{M}, \tilde{G}\right)$ of $\mathbf{M}$ with respect to $\tilde{G}$ we define a dataset of arbitrary structure representing features of $\tilde{G}$ in a way proper for analyzing.

2.2 An Overview of Goal Representations

2.2.1 Goal Signals

The goal representation which is frequently used in the literature is the *goal signal* $\theta = \left(\theta_1, \theta_2, \ldots, \theta_{l_p}\right)$, which reflects such potentially sensitive properties of a group as [4, p. 77] extreme values, statistical features, etc. For the sake of simplicity, we assume that each goal signal value corresponds to one parameter submicrofile $\mathbf{M}_k$, $k = \overline{1, l_p}$. The goal signal may be treated as a function $\theta = \theta(\mathbf{P}, \mathbf{V})$ of parameter values $\mathbf{P}$ and a term $\mathbf{V}$ defining the set of vital value combinations, with each $\theta_k = \theta(P_k, \mathbf{V})$.

In the literature, there are distinguished several kinds of goal signals. Among the more popular ones is the *quantity signal* $\mathbf{q} = \left(q_1, q_2, \ldots, q_{l_p}\right)$ introduced in [14]. The elements q_k, $k = \overline{1, l_p}$, stand for the quantities of respondents with a particular parameter value P_k and values of vital attributes belonging to $\mathbf{V}$.

In many cases, absolute quantities are not representative, and should be replaced with the relative ratios. In these cases, the *concentration signal* $\mathbf{c} = \left(c_1, c_2, \ldots, c_{l_p}\right)$ introduced in [19] is used instead of the quantity one. The elements c_k, $k = \overline{1, l_p}$, are obtained by dividing q_k by the overall number of respondents in a specified parameter submicrofile:

$$c_k = \frac{q_k}{\mu_k}, \quad k = \overline{1, l_p}\ . \tag{1}$$

Vital attributes enable us to split each parameter submicrofile $\mathbf{M}_j$ into *vital submicrofiles* $\mathbf{M}_k^{(G)}$, $k = \overline{1, l_p}$, which contain all microfile records with a parameter value P_k and values of vital attributes belonging to $\mathbf{V}$, and *non-vital submicrofiles* $\mathbf{M}_k^{(\overline{G})}$, $k = \overline{1, l_p}$, which contain the microfile records with a parameter value P_k and values of vital attributes not belonging to $\mathbf{V}$. Each submicrofile $\mathbf{M}_k^{(G)}$ contains q_k records, each submicrofile $\mathbf{M}_k^{(\overline{G})}$ contains $(\mu_k - q_k)$ records.

2.2.2 Goal Surfaces

When we need to deal with the anonymity of fuzzy groups, the goal signal is not sufficient to embrace all the information about the microfile respondents. We need to introduce the generalization of the goal signal called the *goal surface* Θ. It can be treated as a function $\Theta = \Theta\left(\mathbf{P}, \tilde{\mathbf{M}}_{\tilde{G}}, \tilde{\mathbf{V}}\right)$ of parameter values $\mathbf{P}$, membership grades of a particular respondent in the fuzzy group $\tilde{\mathbf{M}}_{\tilde{G}}$, and a term $\tilde{\mathbf{V}}$ defining the set of generalized vital value combinations, with each $\Theta_{jk} = \Theta\left(P_k, \mu_{\tilde{G}j}, \tilde{\mathbf{V}}\right)$.

There can be distinguished two kinds of goal surfaces, a *quantity surface* $\mathbf{Q}$ and a *concentration surface* $\mathbf{C}$. To build $\mathbf{Q}$, one needs to calculate the membership grades $\mu_{\tilde{G}}(u_i)$ in the fuzzy group $\tilde{G}$ for every microfile respondent $u_i \in \tilde{G}$, that is, every respondent whose vital attribute values belong to the universes of discourse of appropriate linguistic variables. This can be carried out by applying a properly designed fuzzy inference system (FIS). In this chapter, we will use the Mamdani FIS [20], which typically consists of several input and output variables, the fuzzification module, the fuzzy inference engine, the fuzzy rule base, and the defuzzification module [21]. Each rule j in the rule base is in the form

$$\text{if } x_1 \text{ is } A_{1j}, \ldots, x_n \text{ is } A_{nj},\ \text{then } y \text{ is } B_j,$$

where $x_1, \ldots, x_n$, and y are input and output variables, respectively; $A_{1j}, \ldots, A_{nj}$, and B_j are values (fuzzy sets) of input and output variables. The inference engine works on the basis of compositional rule of inference (CRI) [22].

To build a FIS, one needs to accomplish the following steps:

1. Choose input, output variables, define appropriate membership functions (MFs).
2. Construct the fuzzy rule base.
3. Choose methods of fuzzy intersection, implication, and aggregation.
4. Choose appropriate defuzzification algorithm.

For the FIS for building **Q**, one needs to take linguistic variables L_j as the input ones. The output variable should represent the membership grade in $\tilde{G}$ of a certain respondent. The generalized vital value combinations should represent the antecedents of the fuzzy rules. In some cases, the problem at hand can impose crisp restrictions to be considered aside from those in the fuzzy rule base.

Having calculated membership grades using FIS, one needs to count the number of respondents with particular parameter values and membership grades in $\tilde{G}$:

$$Q_{jk} = \left| \left\{ u_i \left| z_{iw_p} = P_k, \mu_{\tilde{G}}(u_i) = \mu_{\tilde{G}j} \right. \right\} \right| . \tag{2}$$

However, in most cases exact values of membership grades are irrelevant and do not shed much light on the distribution to be analyzed. The numbers of membership grades belonging to certain intervals can provide information that is much more useful. We need to split $\tilde{\mathbf{M}}_{\tilde{G}}$ into intervals $\Delta_{\tilde{\mathbf{M}}s}$, $s = \overline{1, r}$, and count the number of respondents with membership grades belonging to them:

$$Q_{sk} = \left| \left\{ u_i \left| z_{iw_p} = P_k, \mu_{\tilde{G}}(u_i) \in \Delta_{\tilde{\mathbf{M}}s} \right. \right\} \right| . \tag{3}$$

It is wise to build (3) only for intervals with high values of membership grades.

Using (3), each parameter submicrofile $\mathbf{M}_k$ may be split into *vital submicrofiles of grade* $\Delta_{\tilde{\mathbf{M}}s}$ $\mathbf{M}_k^{\left(\tilde{G}_{\Delta_{\tilde{\mathbf{M}}s}}\right)}$ and *non-vital submicrofiles* $\mathbf{M}_k^{\left(\tilde{G}_{\Delta_{\tilde{\mathbf{M}}0}}\right)}$, $k = \overline{1, l_p}$, $s = \overline{1, r}$. Vital submicrofiles contain the microfile records u_i with the parameter value P_k and $\mu_{\tilde{G}}(u_i) \in \Delta_{\tilde{\mathbf{M}}s}$. Non-vital submicrofiles contain the microfile records u_i with the parameter value P_k and the membership grade in $\tilde{G}$ not belonging to any interval. Each vital submicrofile of a certain grade $\mathbf{M}_k^{\left(\tilde{G}_{\Delta_{\tilde{\mathbf{M}}s}}\right)}$ contains Q_{sk} records, each non-vital submicrofile $\mathbf{M}_k^{\left(\tilde{G}_{\Delta_{\tilde{\mathbf{M}}0}}\right)}$ contains $\left(\mu_k - \sum_{s=1}^{r} Q_{sk}\right)$ records.

In cases when the absolute numbers of respondents are not representative, it is better to use the concentration surface **C** with the elements

$$C_{sk} = \frac{Q_{sk}}{\mu_k}, \quad k = \overline{1, l_p} . \tag{4}$$

2.3 The General Approach to Creating the FIS for Violating Anonymity of a Fuzzy Group

In Sect. 2.2.2, we outlined several steps that need to be accomplished to create a FIS for building a quantity surface **Q** for a fuzzy respondent group $\tilde{G}$. However, when the task of violating anonymity of a fuzzy group is concerned, the group whose anonymity needs to be violated is not precisely defined. For example, when the task is to violate anonymity of a fuzzy group of "respondents who can be considered military with the high level of confidence," it is not clear what vital attributes should be taken, and what values the corresponding linguistic variables have.

In general, to build a FIS for classifying respondents as belonging to a given fuzzy group with a certain grade, one needs to proceed according to such steps:

1. According to external statistical data and/or expert judgment, determine the microfile attributes, which can be used in combination to describe respondents belonging to the fuzzy group with a *high* membership grade.
2. Split, if necessary, the values of these attributes into meaningful intervals, and obtain the distributions over the values of each attribute for the respondents belonging to the fuzzy group with a high membership grade.
3. Define the ranges of the values of these attributes, outside which respondents are considered (in a crisp way) as not belonging to the group.
4. Exclude from the set of the attributes defined on step 1 those ones, distribution over which is sufficiently close to the uniform one.
5. Exclude from the set of the attributes defined on step 4 those ones, distribution over which for the respondents belonging to the fuzzy group with a high membership grade is sufficiently close to the distribution for the respondents at large.
6. According to external statistical data and/or expert judgment, determine the microfile attributes which can be used in combination to describe respondents belonging to the fuzzy group with a *low* membership grade, and add them to the set defined on step 5.
7. Split, if necessary, the values of the newly added attributes into meaningful intervals, and obtain the distributions over the values of each attribute for the respondents belonging to the fuzzy group with a low membership grade.
8. Define the ranges of the values of the newly added attributes inside which respondents are considered (in a crisp way) as not belonging to the group.
9. Define the values of all the input variables of the FIS. Variables correspond to some or all of the attributes from the set defined on step 6, 10. Judging from external statistical data and/or expert judgment, define values of the output linguistic variable, and construct a meaningful set of fuzzy rules.
10. Choose appropriate methods of fuzzy union, intersection, implication, aggregation, and defuzzification.

3 Practical Results

3.1 Violating Anonymity of the Crisp Group of Military Personnel

In this section, we will show how anonymity of the crisp group of respondents can be violated using publicly available microfile data. In particular, we want to show how the potential sites of the military bases can be determined using the regional distribution of the military personnel. For our purpose, we used the 5-Percent Public

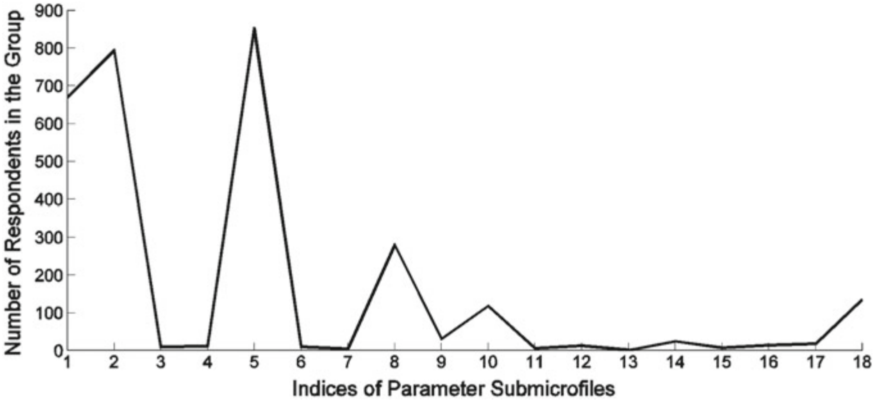

Fig. 2 The quantity signal obtained for the crisp group of active duty military personnel

Use Microdata Sample Files from the U. S. Census Bureau [23] corresponding to the 2000 U. S. Census microfile data on the state of Florida.

In accordance with Sect. 2.1, we took "Place of Work Super-PUMA" (where PUMA stands for "Public Use Microdata Area") as the parameter attribute. We took codes of all the statistical areas of Florida, i. e. each 10th value in the range 12010–12180, as the parameter values. With the help of these parameter values, the microfile can be split into 18 parameter submicrofiles, $\mathbf{M}_1, \ldots, \mathbf{M}_{18}$, with the total number of respondents in each of them, $\mu_1, \ldots, \mu_{18}$, given as follows:

$$\upmu = (\mu_1, \ldots, \mu_{18}) = (8375, 10759, 9683, 10860, 25753, 10153, 6916, 50680, \\ 39892, 10453, 9392, 9016, 8784, 11523, 11158, 24124, 30666, 46177)\ . \tag{5}$$

We took "Military Service" as a vital attribute. Its value "1," standing for "Active Duty," was chosen as the only vital value. Thus, we have defined the *group G* of active duty military personnel distributed over statistical areas of Florida. The quantity signal $\mathbf{q}$ is shown in Fig. 2.

As we see, there are three extreme values in the quantity signal. More precisely, above 75 % of all the active duty military personnel work in the first, second, and fifth statistical areas. Such disproportionate quantities may point to the sites of military bases. Thus, anonymity can be violated relatively easily for a crisp group.

3.2 *Violating Anonymity of the Fuzzy Group of Military Enlisted Members*

In the previous section, we showed that anonymity of a crisp group of military personnel can be violated by analyzing extreme values of an appropriate quantity signal. One of the crudest ways to prevent such violation is to remove completely from the microfile the "Military Service" attribute. However, as we show in this section, anonymity can also be violated for a fuzzy group $\tilde{G}$ of "respondents who can be considered military enlisted members with the high level of confidence."

To construct a quantity surface, we need to build appropriate FIS. We decided to use the demographic analysis of the military personnel conducted by the Office of the Deputy under Secretary of Defense [24] in 2011 and updated in November 2012 as our main source of relevant statistical data. We also used certain expert judgments, e.g. that the military enlisted members in majority tend to work more than 40 h per week. We then followed along the steps outlined in Sect. 2.3:

1. We chose microfile attributes "Age," "Sex," "Black or African American," "Marital Status," "Educational Attainment," and "Hours per Week in 1999" as the ones that can be used in combination to describe respondents belonging to our fuzzy group with a high membership grade.
2. According to [24], the distributions of the active duty enlisted members over the values of the chosen attributes are as follows:
   - 49.3 % are 25 years of age or younger, 22.8 % are 26–30 years of age, 13.1 % are 31 to 35 years of age, 9.2 % are 36–40 years of age, 5.5 % are 41 years of age or older;
   - 85.8 % are male, and 14.2 % are female;
   - 16.9 % are Black or African American, whereas 83.1 % are not;
   - 54.0 % are married, 41.3 % never married, and 4.6 % are divorced;
   - 93.4 % have less than Bachelor's Degree, 5.3 % have Bachelor's or Advanced Degree (other 1.3 % either have no High School diploma, or their educational level is unknown).
3. Having analyzed information presented in [24], we decided to consider respondents whose are younger than 18 years of age or older than 45 years of age as those ones who do not belong to our fuzzy group in a crisp sense.
4. We excluded from the set of supposedly vital attributes "Marital Status" because it provides the distribution, which is very close to the uniform one.
5. We decided to skip this step since all attributes provide significant information.
6. Using expert judgment that every enlisted member has to exhibit a certain level of English, we added the attribute "English Ability" to our set of attributes.
7. We decided to skip this step as not necessary.
8. We decided to choose "English Ability" values "3" and "4" (standing for "Not well" and "Not at all," respectively) as those ones which correspond to respondents who do not belong to the fuzzy group in the crisp sense.
9. We decided to take five input variables for the FIS, namely, "Age," "Sex," "Black or African American," "Educational Attainment," and "Hours per Week." Values of "Age," "Sex," and "Hours per Week" are presented in Figs. 3–5, respectively (codes for the "Educational Attainment" variable are given in Table 1). Variable "Sex" has two values, "Male" and "Female," with the MFs

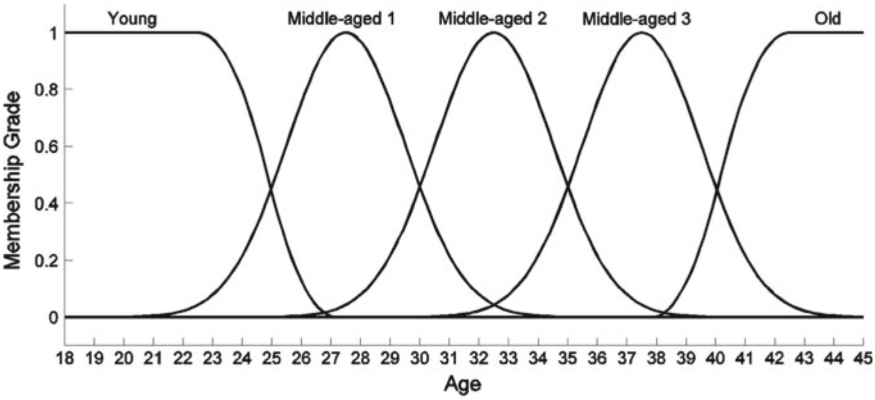

Fig. 3 Membership functions for the "Age" variable

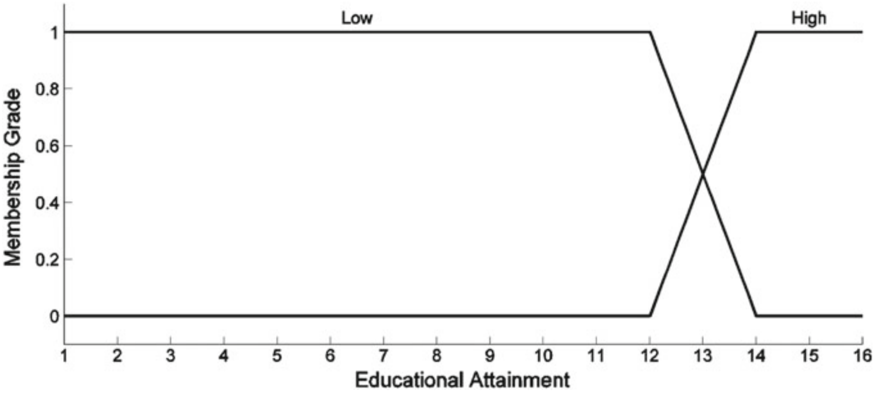

Fig. 4 Membership functions for the "Educational Attainment" variable

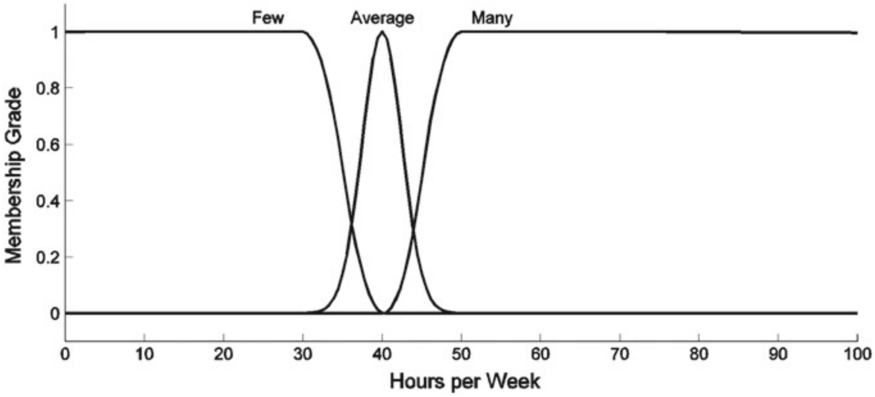

Fig. 5 Membership functions for the "Hours per Week" variable

Table 1 Codes for the "Educational Attainment" variable

Code	Description	Code	Description
1	No schooling completed	9	High school graduate
2	Nursery school to 4th grade	10	Some college, but less than 1 year
3	5th grade or 6th grade	11	One or more years of college, no degree
4	7th grade or 8th grade	12	Associate degree
5	9th grade	13	Bachelor's degree
6	10th grade	14	Master's degree
7	11th grade	15	Professional degree
8	12th grade, no diploma	16	Doctorate degree

$$\mu_{\text{Male}}(x) = \begin{cases} 1, & x = 1 \\ 0, & x \neq 1 \end{cases}, \quad \mu_{\text{Female}}(x) = \begin{cases} 1, & x = 2 \\ 0, & x \neq 2 \end{cases},$$

where "1" is the microfile attribute value standing for "Male," and "2" is the value standing for "Female." Variable "Black or African American" has two values, "No" and "Yes," with the MFs

$$\mu_{\text{No}}(x) = \begin{cases} 1, & x = 0 \\ 0, & x \neq 0 \end{cases}, \quad \mu_{\text{Yes}}(x) = \begin{cases} 1, & x = 1 \\ 0, & x \neq 1 \end{cases},$$

where "0" is the value standing for "Not Black," and "1" is the value standing for "Black."

10. Values of the output variable "Membership in a Fuzzy Group" are presented in Fig. 6. The set of rules is presented in Table 2. These rules were largely determined by analyzing [24]. For instance, if almost half of all enlisted members are young, many of them work more than 40 h per week, and the absolute majority are "White," "Male," and "Lowly educated," then respondents with such characteristics can be considered enlisted members with "high" membership grade. For less obvious vital value combinations we used expert judgment.
11. We decided to take *maximum* as fuzzy union and aggregation, *minimum* as fuzzy intersection and implication, and *centroid method* for defuzzification.

Using the FIS constructed in accordance with these 11 steps, we calculated membership grades for all the respondents in the microfile that belong to the group in a crisp sense. We decided to choose the following intervals to construct the quantity surface **Q** (3): $\Delta_{\tilde{\mathbf{M}}1} = (0.5; 0.6]$, $\Delta_{\tilde{\mathbf{M}}2} = (0.6; 0.7]$, $\Delta_{\tilde{\mathbf{M}}3} = (0.7; 0.8]$, $\Delta_{\tilde{\mathbf{M}}4} = (0.8; 0.9]$.

The quantity surface does not provide necessary information for violating anonymity of the fuzzy group. To determine potential sites of military bases, it is better to use the concentration surface **C** (4) obtained using (5). Surfaces **Q** and **C** are given below (we present all the results with three decimal numbers; the calculations throughout the chapter had been carried out with higher precision):

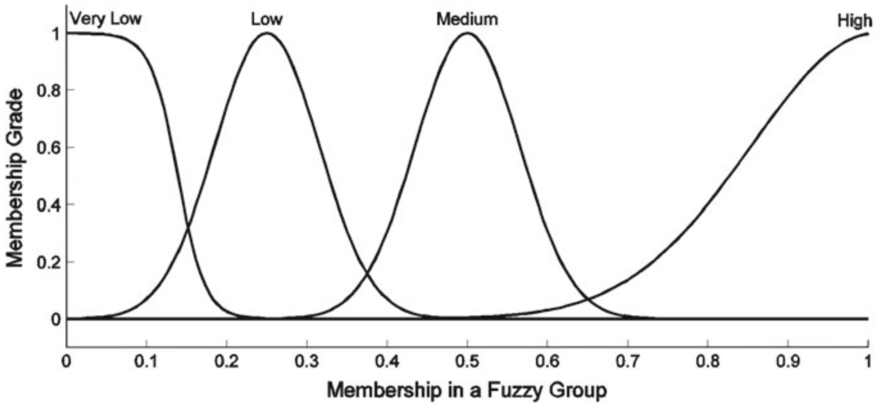

Fig. 6 Membership functions for the “Membership in a Fuzzy Group” variable

$$\mathbf{Q}^T = \begin{pmatrix} 204 & 23 & 56 & 328 \\ 218 & 54 & 94 & 377 \\ 159 & 12 & 46 & 183 \\ 179 & 29 & 53 & 284 \\ 438 & 97 & 151 & 730 \\ 160 & 28 & 41 & 211 \\ 116 & 25 & 34 & 170 \\ 834 & 159 & 238 & 1099 \\ 745 & 142 & 226 & 900 \\ 144 & 23 & 45 & 192 \\ 155 & 35 & 49 & 276 \\ 144 & 31 & 48 & 191 \\ 150 & 33 & 46 & 183 \\ 194 & 30 & 53 & 245 \\ 176 & 31 & 60 & 238 \\ 330 & 57 & 95 & 374 \\ 433 & 62 & 122 & 439 \\ 619 & 86 & 192 & 738 \end{pmatrix}, \quad \mathbf{C}^T = \begin{pmatrix} 0.024 & 0.003 & 0.007 & 0.039 \\ 0.020 & 0.005 & 0.009 & 0.035 \\ 0.016 & 0.001 & 0.005 & 0.019 \\ 0.016 & 0.003 & 0.005 & 0.026 \\ 0.017 & 0.004 & 0.006 & 0.028 \\ 0.016 & 0.003 & 0.004 & 0.021 \\ 0.017 & 0.004 & 0.005 & 0.025 \\ 0.016 & 0.003 & 0.005 & 0.022 \\ 0.019 & 0.004 & 0.006 & 0.023 \\ 0.014 & 0.002 & 0.004 & 0.018 \\ 0.017 & 0.004 & 0.005 & 0.029 \\ 0.016 & 0.003 & 0.005 & 0.021 \\ 0.017 & 0.004 & 0.005 & 0.021 \\ 0.017 & 0.003 & 0.005 & 0.021 \\ 0.016 & 0.003 & 0.005 & 0.021 \\ 0.014 & 0.002 & 0.004 & 0.016 \\ 0.014 & 0.002 & 0.004 & 0.014 \\ 0.013 & 0.002 & 0.004 & 0.016 \end{pmatrix}.$$

The sum of rows of **C** is shown in Fig. 7 along with the superimposed quantity signal **q** obtained in Sect. 3.1 (to fit the scale, we normalized both vectors by dividing them by their maximal values). By analyzing extreme values obtained from the concentration surface **C**, we can determine the same statistical areas we determined in Sect. 3.1. It is worth noting that extreme value in the element 11 was not present in **q**, however, all the extremes that actually *were* present in **q** have been successfully determined, even though the attribute “Military Service” was removed from the microfile.

Thus, we successfully managed to violate anonymity for the fuzzy group of “respondents who can be considered military enlisted members with the high level

Table 2 Fuzzy rule base for the FIS in example

Hours per week	Educational attainment	Sex	Black or Afr. Amer.	Age				
				Young	Mid. -aged 1	Mid. -aged 2	Mid. -aged 3	Old
Few	Low	Male	Yes	VL	VL	VL	VL	VL
			No	L	VL	VL	VL	VL
		Female	Yes	VL	VL	VL	VL	VL
			No	VL	VL	VL	VL	VL
	High	Male	Yes	VL	VL	VL	VL	VL
			No	VL	VL	VL	VL	VL
		Female	Yes	VL	VL	VL	VL	VL
			No	VL	VL	VL	VL	VL
	Low	Male	Yes	L	VL	VL	VL	VL
Average			No	H	L	L	L	L
		Female	Yes	VL	VL	VL	VL	VL
			No	L	VL	VL	VL	VL
	High	Male	Yes	VL	VL	VL	VL	VL
			No	VL	VL	VL	VL	VL
		Female	Yes	VL	VL	VL	VL	VL
			No	VL	VL	VL	VL	VL
	Low	Male	Yes	L	VL	VL	VL	VL
			No	H	M	M	M	L
Many		Female	Yes	VL	VL	VL	VL	VL
			No	L	VL	VL	VL	VL
	High	Male	Yes	VL	VL	VL	VL	VL
			No	VL	L	L	L	L
		Female	Yes	VL	VL	VL	VL	VL
			No	VL	VL	VL	VL	VL

of confidence." In other words, even if group anonymity is provided for a crisp group of military personnel (Sect. 3.1), it is still possible to retrieve sensitive information from the microfile using the concept of a fuzzy respondent group.

4 Providing Anonymity for Crisp and Fuzzy Respondent Groups

4.1 The Generic Scheme of Providing Group Anonymity

The *task of providing group anonymity* in a microfile is the task of modifying it for a group $\tilde{G}(\tilde{\mathbf{V}}, \mathbf{P})$, so that sensitive (for the task solved) data features become confided. The generic scheme of providing group anonymity goes as follows:

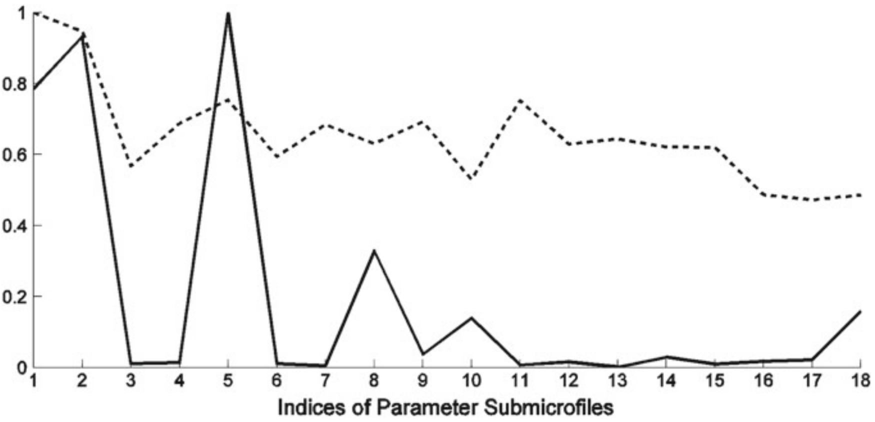

Fig. 7 The quantity signal (*solid line*) and the sum of the rows of the concentration surface (*dashed line*) for the example

1. Prepare a depersonalized microfile **M**.
2. Define groups $\tilde{G}_i(\tilde{\mathbf{V}}_i, \mathbf{P})$, $i = \overline{1, k}$, representing respondents to be protected.
3. For each i from 1 to k:

   - choose data *goal representation* $\Omega_i(\mathbf{M}, \tilde{G}_i)$ representing particular features of the group in a way appropriate for its further modification;
   - define the *goal mapping function* $\Upsilon_i : \mathbf{M} \to \Omega_i(\mathbf{M}, \tilde{G}_i)$ and obtain the goal representation;
   - define the *modifying functional* $\Xi_i : \Omega_i(\mathbf{M}, \tilde{G}_i) \to \Omega_i^*(\mathbf{M}, \tilde{G}_i)$ and obtain the *modified goal representation*;
   - define the *inverse goal mapping function* $\Upsilon_i^{-1} : \Omega_i^*(\mathbf{M}, \tilde{G}_i) \to \mathbf{M}^*$ and obtain the modified microfile.

4. Prepare the modified microfile $\mathbf{M}^*$ for publishing.

The first three operations at step 3 constitute the *first stage* of solving the TPGA. Obtaining the modified microfile using the inverse goal mapping function at step 3 is the only operation constituting the *second stage* of solving the TPGA.

4.2 Wavelet Transforms as the Modifying Functional

4.2.1 One-Dimensional Wavelet Transforms as the Modifying Functional for the Goal Signals

We will introduce wavelet transforms (WT) to the extent necessary for applying them to modifying the goal signal. For more information on wavelets, consult [25]. For a detailed discussion of applying WT to solving the TPGA, refer to [4].

Let $\mathbf{h} = (h_1, h_2, \dots, h_n)$ and $\mathbf{l} = (l_1, l_2, \dots, l_n)$ denote the *high-frequency* and *low-frequency wavelet filter*, respectively. To perform the goal signal one-level wavelet decomposition, we need to perform the following operations:

$$\mathbf{a}_1 = \theta *_{\downarrow 2} \mathbf{l} \ , \quad \mathbf{d}_1 = \theta *_{\downarrow 2} \mathbf{h} \ , \tag{6}$$

where $*_{\downarrow 2}$ denotes the convolution with the follow-up dyadic downsampling, array $\mathbf{a}_1$ ($\mathbf{d}_1$) consists of level one approximation (detail) coefficients.

To simplify the notation, let us introduce the following operations:

$$\mathbf{z} = \Big(\big(\theta *_{\downarrow 2} \mathbf{f}\big) \underbrace{*_{\downarrow 2} \mathbf{f}\big) \dots *_{\downarrow 2} \mathbf{f}}_{k-1 \text{ times}} = \prod_{i=1}^{k} \big(\theta *_{\downarrow 2} \mathbf{f}\big) \ , \tag{7}$$

$$\mathbf{z} = \Big(\big(\theta *_{\uparrow 2} \mathbf{f}\big) \underbrace{*_{\uparrow 2} \mathbf{f}\big) \dots *_{\uparrow 2} \mathbf{f}}_{k-1 \text{ times}} = \prod_{i=1}^{k} \big(\theta *_{\uparrow 2} \mathbf{f}\big) \ , \tag{8}$$

where $*_{\uparrow 2}$ denotes the dyadic upsampling with the follow-up convolution.

To obtain decomposition coefficients of arbitrary level k, we need to perform (6) with the goal signal replaced by the approximation coefficients of level $k-1$:

$$\mathbf{a}_k = \prod_{i=1}^{k} \big(\theta *_{\downarrow 2} \mathbf{l}\big) \ , \quad \mathbf{d}_k = \left(\prod_{i=1}^{k-1} \big(\theta *_{\downarrow 2} \mathbf{l}\big)\right) *_{\downarrow 2} \mathbf{h} \ . \tag{9}$$

To obtain the goal signal approximation and details of level k, we need to perform the following operations:

$$\mathbf{A}_k = \prod_{i=1}^{k} \big(\mathbf{a}_k *_{\uparrow 2} \mathbf{l}\big) \ , \quad \mathbf{D}_k = \prod_{i=1}^{k-1} \Big(\big(\mathbf{d}_k *_{\uparrow 2} \mathbf{h}\big) *_{\uparrow 2} \mathbf{l}\Big) \ . \tag{10}$$

The goal signal can be decomposed into the following sum:

$$\theta = \mathbf{A}_k + \sum_{i=1}^{k} \mathbf{D}_i \ . \tag{11}$$

Wavelet approximation $\mathbf{A}_k$ of the signal represents its smoothed version. Wavelet details of all levels $\mathbf{D}_i$, $i = \overline{1,k}$, represent high-frequency fluctuations in it.

To protect such properties of the goal signal as its extreme values, two different approaches may be proposed [4]. According to the *extremum transition approach*, the goal signal has to be modified in such way that its new extreme values differ from the initial ones. The other approach called the *Ali Baba's wife approach* implies not eliminating existing extreme values but adding several new alleged ones, which makes it impossible to discriminate between real and fake extreme values.

Aside from protecting signal properties, it is important to guarantee that the overall data utility is not reduced very much. WT can be successfully applied in order to achieve both goals. To mask extreme values, we can modify the goal signal approximation, whereas leaving the signal details intact (or modifying them at most proportionally) preserves important properties of the initial data.

However, mere modifying the approximation will not do much good, because internal structure of the signal will be tampered with. The better way of modifying the signal approximation is to modify its approximation coefficients. To do this, we need to know the explicit dependence of the approximation values on the approximation coefficients. This dependence can be retrieved from the so called wavelet reconstruction matrix (WRM) $\mathbf{M}_{rec}$ introduced in [14]:

$$\mathbf{A}_k = \mathbf{M}_{rec} \cdot \mathbf{a}_k \ . \tag{12}$$

With the help of the WRM, we can represent each approximation element as the linear combination of approximation coefficients and $\mathbf{M}_{rec}$ elements. The latter ones are dependent on wavelet filter elements and the size of the goal signal. Using (12), we can construct restrictions for the linear programming problem, whose solution yields modified approximation coefficients $\tilde{\mathbf{a}}_k$. These coefficients can be used to obtain modified approximation $\tilde{\mathbf{A}}_k$ according to (10).

Using (11), we can obtain the signal $\breve{\theta} = \tilde{\mathbf{A}}_k + \sum_{i=1}^{k} \mathbf{D}_i$. If any of its elements are negative, we need to add to the signal a sufficiently great number γ to make all the signal entries non-negative. To preserve the mean value of the goal signal after this operation, we need to multiply it by an appropriate coefficient:

$$\theta^* = \left(\breve{\theta} + \gamma\right) \cdot \frac{\sum\limits_{k=1}^{l_p} \theta_k}{\sum\limits_{k=1}^{l_p} \left(\breve{\theta}_k + \gamma\right)} \ . \tag{13}$$

When the goal signal is the concentration signal, it is necessary to apply (13) not only to the signal itself, but to the corresponding quantity signal as well, so that the overall number of respondents in the microfile does not change.

4.2.2 Separable Two-Dimensional Wavelet Transforms as the Modifying Functional for the Goal Surfaces

To perform one-level wavelet decomposition of the goal surface Θ, we need to carry out the following calculations:

$$\begin{aligned}\mathbf{a}_1 &= \overbrace{\underbrace{\left(\Theta *_{\downarrow 2} \mathbf{l}\right)}_{\text{row}-\text{wise}} *_{\downarrow 2} \mathbf{l}}^{\text{column}-\text{wise}} \,, \mathbf{d}_{h1} = \overbrace{\underbrace{\left(\Theta *_{\downarrow 2} \mathbf{l}\right)}_{\text{row}-\text{wise}} *_{\downarrow 2} \mathbf{h}}^{\text{column}-\text{wise}} \,, \\ \mathbf{d}_{v1} &= \overbrace{\underbrace{\left(\Theta *_{\downarrow 2} \mathbf{h}\right)}_{\text{row}-\text{wise}} *_{\downarrow 2} \mathbf{l}}^{\text{column}-\text{wise}} \,, \mathbf{d}_{d1} = \overbrace{\underbrace{\left(\Theta *_{\downarrow 2} \mathbf{h}\right)}_{\text{row}-\text{wise}} *_{\downarrow 2} \mathbf{h}}^{\text{column}-\text{wise}} \,. \end{aligned} \tag{14}$$

These operations are the generalized versions of (6). However, instead of one array of detail coefficients, we obtain three of them, i.e. *horizontal detail coefficients* $\mathbf{d}_{h1}$, *vertical detail coefficients* $\mathbf{d}_{v1}$, and *diagonal detail coefficients* $\mathbf{d}_{d1}$.

The goal surface can be decomposed into the sum of its approximation and three types of details:

$$\Theta = \mathbf{A}_k + \sum_{i=1}^{k} \mathbf{D}_{hi} + \sum_{i=1}^{k} \mathbf{D}_{vi} + \sum_{i=1}^{k} \mathbf{D}_{di} \ . \tag{15}$$

To modify the goal surface using WT, we can use the method similar to the one described in Sect. 4.2.1. Each element of the two-dimensional approximation can be presented as the linear combination of the approximation coefficients and some values dependent on the wavelet filter elements and the size of the goal surface. This representation is useful for constructing restrictions of a linear programming problem, whose solution yields modified approximation coefficients $\tilde{\mathbf{a}}_k$.

Applying (15), we can obtain the surface $\breve{\Theta} = \tilde{\mathbf{A}}_k + \sum_{i=1}^{k} \mathbf{D}_{hi} + \sum_{i=1}^{k} \mathbf{D}_{vi} + \sum_{i=1}^{k} \mathbf{D}_{di}$, which can be amended if necessary using the procedure described in Sect. 4.2.1 yielding the modified goal surface Θ^*:

$$\Theta^* = \left(\breve{\Theta} + \gamma\right) \cdot \frac{\sum\limits_{s=1}^{r} \sum\limits_{k=1}^{l_p} \Theta_{sk}}{\sum\limits_{s=1}^{r} \sum\limits_{k=1}^{l_p} \left(\breve{\Theta}_{sk} + \gamma\right)} \ . \tag{16}$$

When the goal surface is the concentration surface, it is necessary to apply (16) not only to the surface itself, but to the corresponding quantity surface as well. In the latter case, the surface needs to be rounded afterwards. If the sum of all the surface elements differs from the initial one after such rounding by a small number ϵ, it is permissible to add ϵ to the greatest element of the rounded surface.

4.3 *Inverse Goal Mapping Functions for Minimizing Microfile Distortion*

4.3.1 Inverse Goal Mapping Functions for Crisp Respondent Groups

Modifying the microfile in order to adjust it to the modified goal representation by applying inverse goal mapping function implies introducing into it a certain level of distortion, whose overall amount has to be minimized. In general, it is a good practice to modify the microfile by applying the inverse goal mapping function to the modified quantity signal (or surface), even when the goal representation is the concentration signal (or surface). In this section, we will assume that the inverse goal mapping function is applied to the modified quantity signal $\mathbf{q}^*$.

To modify the microfile in order to adjust it to the modified quantity signal $\mathbf{q}^*$, one needs to alter values of the parameter attribute for certain respondents. To make sure that the number of respondents in each parameter submicrofile remains the same, respondents should be altered in pairs. One of the respondents in a pair has to belong to the group G, whereas the other one has to lie outside the group. We call this operation the *swapping of the respondents between the submicrofiles* (SRBS).

Let *influential attributes* [3] be the ones, whose distribution plays a great role for researchers. To minimize overall microfile distortion, one needs to search for pairs of respondents to swap between submicrofiles that are close to each other. To determine how "close" respondents are, one can use the *influential metric* [3]:

$$\mathrm{InfM}\left(u_1, u_2\right) = \sum_{l=1}^{n_{ord}} \omega_l \left(\frac{u_1\left(I_l\right) - u_2\left(I_l\right)}{u_1\left(I_l\right) + u_2\left(I_l\right)} \right)^2 + \sum_{k=1}^{n_{nom}} \gamma_k \chi^2 \left(u_1\left(J_k\right), u_2\left(J_k\right)\right) \;, \quad (17)$$

where I_l stands for the lth ordinal influential attribute (their total number is n_{ord}); J_k stands for the kth nominal influential attribute (their total number is n_{nom}); $u\left(\cdot\right)$ returns respondent u's specified attribute value; $\chi\left(v_1, v_2\right)$ is equal to χ_1 if values v_1 and v_2 fall into one category, and χ_2 otherwise; ω_l and γ_k are non-negative weighting coefficients to be taken judging from the importance of a certain attribute (the more important is the attribute, the greater is the coefficient).

To organize the process of the pairwise SRBS, let us introduce the notion of the *valence* δ_k^i of the submicrofile $\mathbf{M}_k^i$ as a number, whose absolute value determines how many respondents need to be added to or removed from the submicrofile, and whose sign shows whether the respondents need to be added (negative valence) or removed (positive valence) from the submicrofile. The valences of the vital submicrofiles $\mathbf{M}_k^{(G)}$, $k = \overline{1, l_p}$, are equal to the values of the so called *difference signal*

$$\delta^{(G)} = \mathbf{q} - \mathbf{q}^* \,. \quad (18)$$

The valences of the non-vital submicrofiles $\mathbf{M}_k^{\overline{(G)}}$, $k = \overline{1, l_p}$, are determined to ensure that the number of respondents in each parameter submicrofile is the same:

Table 3 The valence matrix for anonymizing crisp respondent groups

	P_1	P_2	$\dots$	P_{l_p}
G	$\delta_1^{(G)}$	$\delta_2^{(G)}$	$\dots$	$\delta_{l_p}^{(G)}$
$\overline{G}$	$\delta_1^{(\overline{G})}$	$\delta_2^{(\overline{G})}$	$\dots$	$\delta_{l_p}^{(\overline{G})}$

$$\delta^{(\overline{G})} = (\mu_k - q_k) - \left(\mu_k - q_k^*\right) = -\,\delta^{(G)}\,, \quad k = \overline{1, l_p}\,. \tag{19}$$

Valences of submicrofiles can be arranged into the *valence matrix* Δ (Table 3).

Performing the swapping is expressed with the help of the *swapping cycle*:

$$C = ((i_1, j_1)\,, (i_1, j_2)\,, (i_2, j_2)\,, (i_2, j_1))\,, \tag{20}$$

where (i_1, j_1) determines the positive valence of the vital submicrofile: $\Delta_{i_1 j_1} > 0$, $i_1 = 1$; (i_1, j_2) determines the negative valence of the vital submicrofile: $\Delta_{i_1 j_2} < 0$, $i_1 = 1$; (i_2, j_2) determines the positive valence of the non-vital submicrofile: $\Delta_{i_2 j_2} > 0$, $i_2 = 2$; (i_2, j_1) determines the negative valence of the non-vital submicrofile: $\Delta_{i_2 j_1} < 0$, $i_2 = 2$; $i_1 \neq i_2$, $j_1 \neq j_2$. Cycle entries are called *cycle vertices*.

To define the swapping cycle, it is sufficient to specify its first two vertices.

Respondents to be swapped over C have to belong to the submicrofiles with positive valences and be close with respect to (17). The *swap* is a triplet

$$S = \langle C, I_1, I_2 \rangle, \tag{21}$$

where C is the cycle (20); I_1 is the index of the respondent (the *first candidate to be swapped*, FCS) in the vital submicrofile with the valence defined by the first C vertex; I_2 is the index of the respondent (the *second candidate to be swapped*, SCS) in the non-vital submicrofile with the valence defined by the third C vertex.

The SRBS over C is interpreted as the transferring of the FCS from the submicrofile defined by the vertex 1 to the one defined by the vertex 2, and the simultaneous transferring of the SCS from the submicrofile defined by the vertex 3 to the one defined by the vertex 4. After performing the SRBS according to S, one needs to reduce by one $\Delta_{i_1 j_1}$ and $\Delta_{i_2 j_2}$, and add one to $\Delta_{i_1 j_2}$ and $\Delta_{i_2 j_1}$.

The *cost* of the swap $c\,(S)$ is a value of (17) calculated for the FCS and SCS. The task of modifying the microfile at the second stage of solving the TPGA lies in determining such an ordered sequence of swaps called the *swapping plan* $\mathbf{S} = \left(S_1, \dots, S_{n_{swap}}\right)$ that satisfies two conditions:

1. After performing all the swaps, $\Delta_{ik} = 0\ \forall i = \overline{1, r}\ \forall k = \overline{1, l_p}$.
2. The overall cost of the swapping plan $c\,(\mathbf{S}) = \sum_{i=1}^{n_{swap}} c\,(S_i)$ has to be minimal.

This task is the one that can be solved using only exhaustive search, so heuristic strategies need to be developed for constructing the swapping plan that yields results acceptable from both the computational complexity and the minimal swap-

Table 4 The valence matrix for anonymizing fuzzy respondent groups

	P_1	P_2	...	P_{l_p}
$\tilde{G}_{\Delta_{\tilde{\mathbf{M}}1}}$	$\delta_1^{\left(\tilde{G}_{\Delta_{\tilde{\mathbf{M}}1}}\right)}$	$\delta_2^{\left(\tilde{G}_{\Delta_{\tilde{\mathbf{M}}1}}\right)}$	...	$\delta_{l_p}^{\left(\tilde{G}_{\Delta_{\tilde{\mathbf{M}}1}}\right)}$
...	...	...	...	...
$\tilde{G}_{\Delta_{\tilde{\mathbf{M}}r}}$	$\delta_1^{\left(\tilde{G}_{\Delta_{\tilde{\mathbf{M}}r}}\right)}$	$\delta_2^{\left(\tilde{G}_{\Delta_{\tilde{\mathbf{M}}r}}\right)}$	...	$\delta_{l_p}^{\left(\tilde{G}_{\Delta_{\tilde{\mathbf{M}}r}}\right)}$
$\tilde{G}_{\Delta_{\tilde{\mathbf{M}}0}}$	$\delta_1^{\left(\tilde{G}_{\Delta_{\tilde{\mathbf{M}}0}}\right)}$	$\delta_2^{\left(\tilde{G}_{\Delta_{\tilde{\mathbf{M}}0}}\right)}$	...	$\delta_{l_p}^{\left(\tilde{G}_{\Delta_{\tilde{\mathbf{M}}0}}\right)}$

ping plan cost points of view. Several strategies that meet these requirements have been proposed in [26].

4.3.2 Inverse Goal Mapping Functions for Fuzzy Respondent Groups

In this section, we will assume that the inverse goal mapping function is applied to the modified quantity surface $\mathbf{Q}^*$.

The valences of the vital submicrofiles of different grades $\mathbf{M}_k^{\left(\tilde{G}_{\Delta_{\tilde{\mathbf{M}}s}}\right)}$, $k = \overline{1, l_p}$, $s = \overline{1, r}$, are equal to the values of the so-called *difference surface*

$$\delta^{(G)} = \mathbf{Q} - \mathbf{Q}^* \,. \tag{22}$$

Valences of non-vital submicrofiles $\mathbf{M}_k^{\left(\tilde{G}_{\Delta_{\tilde{\mathbf{M}}0}}\right)}$, $k = \overline{1, l_p}$, are determined to ensure that the number of respondents in each parameter submicrofile is the same:

$$\delta_k^{\left(\tilde{G}_{\Delta_{\tilde{\mathbf{M}}0}}\right)} = \left(\mu_k - \sum_{i=1}^{r} Q_{ik}\right) - \left(\mu_k - \sum_{i=1}^{r} Q_{ik}^*\right), \quad k = \overline{1, l_p} \,. \tag{23}$$

Valences of submicrofiles can be arranged into the valence matrix Δ (Table 4).

Because of the procedure for obtaining $\mathbf{Q}^*$ using the two-dimensional WT (Sect. 4.2.2), it is impossible to modify the microfile by performing only the SRBS. However, it is possible to modify $\mathbf{M}$ by performing the *transferring of the respondents from one submicrofile* $\mathbf{M}_k^{\left(\tilde{G}_{\Delta_{\tilde{\mathbf{M}}s_1}}\right)}$, $s_1 \geq 0$, *to another* $\mathbf{M}_k^{\left(\tilde{G}_{\Delta_{\tilde{\mathbf{M}}s_2}}\right)}$, $s_2 \geq 0$, $s_1 \neq s_2$. To *transfer* respondent u from $\mathbf{M}_k^{\left(\tilde{G}_{\Delta_{\tilde{\mathbf{M}}s_1}}\right)}$ to $\mathbf{M}_k^{\left(\tilde{G}_{\Delta_{\tilde{\mathbf{M}}s_2}}\right)}$ means to modify its vital attributes values so that its membership grade in $\tilde{G}$ $\mu_{\tilde{G}}(u)$ belongs to the interval $\Delta_{\tilde{\mathbf{M}}s_2}$. The interval $\Delta_{\tilde{\mathbf{M}}0}$ may be viewed as the interval containing all the values from [0, 1], which don't belong to any other interval $\Delta_{\tilde{\mathbf{M}}s}$, $s = \overline{1, r}$.

Performing the transferring is expressed with the help of the *transferring cycle*:

$$C_T = \left(\left(i_1, j_1\right), \left(i_2, j_1\right)\right) \ , \tag{24}$$

where (i_1, j_1) determines the negative valence of the submicrofile: $\varDelta_{i_1 j_1} < 0$; (i_2, j_1) determines the positive valence of the submicrofile: $\varDelta_{i_2 j_1} > 0$; $i_1 \neq i_2$.

We propose to determine the respondent to be transferred in the following way:

1. Randomly choose a respondent from the submicrofile defined by the first cycle vertex (we will call this record the *representative respondent*, RR).
2. Choose the respondent from the submicrofile defined by the second vertex closest to the RR with respect to (17) (we will call this respondent the *candidate to be transferred*, CT).

We will perform the transferring of the CT by equating its vital attribute values to the ones taken from the RR. The *transfer* can be represented as a triplet

$$T = \langle C_T, I_1, I_2 \rangle \ , \tag{25}$$

where C_T is the cycle (24); I_1 is the index of the RR; I_2 is the index of the CT. After performing the transferring according to T, one needs to reduce by one the absolute values of $\varDelta_{i_1 j_1}$ and $\varDelta_{i_2 j_1}$.

The *cost* of the transfer $c\,(T)$ is a value of (17) calculated for the RR and CT. The task of modifying the microfile at the second stage of solving the TPGA can be reduced to determining such an ordered sequence of transfers called the *transferring plan* $\mathbf{T} = \left(T_1, \ldots, T_{n_{trans}}\right)$ that satisfies two conditions:

1. After performing all the transfers, $\varDelta_{ij} = 0 \ \forall i = \overline{1, r} \ \forall j = \overline{1, l_p}$.
2. The overall cost of the transferring plan $c\,(\mathbf{T}) = \sum_{i=1}^{n_{trans}} c\,(T_i)$ has to be minimal.

It is possible to solve the TPGA by performing only transfers, but such approach is not acceptable since it implies perturbing microfile records. We propose to reduce the overall number of the transfers by performing the SRBS beforehand.

The overall number of the transfers to perform in the microfile $\mathbf{M}$ is equal to

$$N_{trans} = \sum_{j=1}^{l_p} \left(\frac{1}{2} \sum_{i=1}^{r} \left| \varDelta_{ij} \right| \right) . \tag{26}$$

After performing the SRBS over the cycle with vertices 1 and 3 corresponding to the positive valences in Δ, and the vertices 2 and 4 corresponding to the negative ones, N_{trans} is reduced by two. Such cycles are called the *full swapping cycles* (FSC). We will denote them by C_F. FSCs are analogous to the ones defined by (20). After performing the SRBS over the cycle with vertices 1, 3, and 4 corresponding to the positive valences in Δ, and the vertex 2 corresponding to the negative one, N_{trans} is reduced by one. Such cycles are called the *partial swapping cycles* (PSC):

$$C_P = \left(\left(i_1, j_1\right), \left(i_1, j_2\right), \left(i_2, j_2\right), \left(i_2, j_1\right)\right) \ , \tag{27}$$

where (i_1, j_1), (i_2, j_2), and (i_2, j_1) determine the positive valences of the submicrofile: $\varDelta_{i_1 j_1} > 0$, $\varDelta_{i_2 j_2} > 0$, $\varDelta_{i_2 j_1} > 0$; (i_1, j_2) determines the negative valence of the submicrofile: $\varDelta_{i_1 j_2} < 0$; $i_1 \neq i_2$, $j_1 \neq j_2$.

To define the swapping cycle, it is sufficient to specify its first three vertices.

Respondents to be swapped over FSC or PSC have to belong to the submicrofiles with the positive valences and be close with respect to (17). The *full swap* can be represented as a triplet

$$S_F = \langle C_F, I_{1F}, I_{2F} \rangle \ , \tag{28}$$

where C_F is the cycle (20); I_{1F} is the index of the respondent (the *first candidate to be fully swapped*, FCFS) from the vital submicrofile with the positive valence defined by the first C_F vertex; I_{2F} is the index of the respondent (the *second candidate to be fully swapped*, SCFS) from the non-vital submicrofile with the positive valence defined by the third C_F vertex.

The *partial swap* can be represented as a triplet

$$S_P = \langle C_P, I_{1P}, I_{2P} \rangle \ , \tag{29}$$

where C_P is the cycle (27); I_{1P} is the index of the respondent (the *first candidate to be partially swapped*, FCPS) from the submicrofile with the positive valence defined by the first C_P vertex; I_{2P} is the index of the respondent (the *second candidate to be partially swapped*, SCPS) from the submicrofile with the positive valence defined by the third C_P vertex.

The *cost* of the swap $c\,(S_F)$ $(c\,(S_P))$ is a value of (17) calculated for the FCFS (FCPS) and SCFS (SCPS) from appropriate submicrofiles. The task of modifying the microfile at the second stage of solving the TPGA for fuzzy respondent groups lies in determining three ordered sequences:

1. The sequence of full swaps called the *full swapping plan* $\mathbf{S}_F = (S_{1F}, \ldots, S_{n_{swapF}})$. The overall cost of the plan $c\,(\mathbf{S}_F) = \sum_{i=1}^{n_{swapF}} c\,(S_{iF})$ has to be minimal. After performing all the swaps from $\mathbf{S}_F$ it is impossible to build full swapping cycles.
2. The sequence of partial swaps called the *partial swapping plan* $\mathbf{S}_P = (S_{1P}, \ldots, S_{n_{swapP}})$. The overall cost of the plan $c\,(\mathbf{S}_P) = \sum_{i=1}^{n_{swapP}} c\,(S_{iP})$ has to be minimal. After performing all the swaps from $\mathbf{S}_P$ it is impossible to build partial swapping cycles.
3. The transferring plan $\mathbf{T} = \left(T_1, \ldots, T_{n_{trans}}\right)$ that has to satisfy two conditions expressed earlier.

The tasks of determining each of three plans are the ones that can be solved using only exhaustive search, so heuristic strategies need to be developed for constructing plans that yield results acceptable from both the computational complexity and the minimal swapping plan cost points of view.

Let $\Delta^{(0)}$ denote the *initial valence matrix*, which is obtained according to Table 4. The generic scheme of all the heuristic strategies for determining the full swapping plan boils down to performing the following steps:

1. Equate the *current valence matrix* to the initial one; set $i = 1$. Perform steps 2–8 while it is possible to build full swapping cycles.
2. Assign $\Delta^{\text{temp}} = \Delta^{(i)}$.
3. By analyzing Δ^{temp}, choose the first vertex of C_{iF}; if it is impossible, stop.
4. Choose the FCFS from the submicrofile defined by the first C_{iF} vertex.
5. By analyzing Δ^{temp}, choose the second vertex of C_{iF}; if it is impossible, equate Δ^{temp} element corresponding to the first C_{iF} vertex to zero and go to 3.
6. By analyzing Δ^{temp}, choose the third vertex of C_{iF}; if it is impossible, equate Δ^{temp} element corresponding to the second C_{iF} vertex to zero and go to 5; otherwise, finish the cycle.
7. Choose the SCFS from the submicrofile defined by the third vertex, which is closest to the first one with respect to (17).
8. Perform the swapping; obtain the current valence matrix $\Delta^{(i)}$ by reducing by one the absolute values of the valences from $\Delta^{(i-1)}$ corresponding to the submicrofiles defined by C_{iF}; set $i = i + 1$; go to 2.

All the strategies differ in particular implementations of steps 3, 4, 5, and 6.

Heuristic strategies for determining the partial swapping plans have the same generic scheme, with several slight differences. Firstly, the first cycle vertex in the case of the partial swapping plans does not necessarily represent the vital microfiles. Secondly, analysis on steps 3, 5, and 6 is carried out using $\Delta^{(i-1)}$, not Δ^{temp}. In addition, the initial valence matrix should be taken as the last current matrix obtained after applying heuristic strategies for determining the full swapping plans.

Since the transferring of the respondents in a parameter submicrofile $\mathbf{M}_k$, $k \in \{1, 2, \ldots, l_p\}$, does not depend on the transferring in any other parameter submicrofile $\mathbf{M}_l$, $l \neq k$, let us discuss the strategies for determining the part of the transferring plan $\mathbf{T}$ corresponding to the kth parameter submicrofile, $k \in \{1, 2, \ldots, l_p\}$.

Let $\Delta^{(0)}_{:k}$ denote the *initial valence matrix column* k, which is the kth column of the valence matrix obtained after performing all swaps. The scheme of the strategies for determining the transferring plan boils down to performing such steps:

1. Equate the *current valence matrix column* k to the initial one; set $i = 1$. Perform steps 2–6 while $\exists l \; \Delta^{(i)}_{lk} \neq \mathbf{0}$.
2. Choose the first vertex of the cycle C_T.
3. Randomly choose the RR from the submicrofile defined by the first C_T vertex.
4. Choose the second vertex of the cycle C_T.
5. Choose the CT closest to the RR with respect to (17).
6. Perform the transferring of the CT; obtain the current valence matrix column k $\Delta^{(i)}_{:k}$ by reducing by one the absolute values of the valences from $\Delta^{(i-1)}_{:k}$ corresponding to the submicrofiles defined by C_T; set $i = i + 1$; go to 2.

All the strategies differ in particular implementations of steps 2 and 4. In this chapter, we decided to use four heuristic strategies for determining swapping cycles by choosing the following implementations of the steps 3, 4, 5, and 6:

1. On step 3, for strategies No. 1 and No. 2 we choose the microfile with the greatest valence, for strategies No. 3 and No. 4—with the smallest one.

2. On step 4, for all strategies we try out all the possible candidates, and choose the one that guarantees the minimum values of (17) on step 8.
3. On step 5, for all strategies we try out all the possible vertices, and choose the one that guarantees the minimum values of (17) on step 8.
4. On step 6, for strategies No. 1 and No. 3 we choose the third vertex from the valence matrix row closest to the row with the first vertex, for strategies No. 2 and No. 4—the third vertex from the last valence matrix row, if possible, or from the row closest to the row with the first vertex, otherwise.

We also chose the following implementations of the steps 2 and 4 of heuristic strategies for determining transferring cycles:

1. On step 2, for strategies No. 1 and No. 2 we choose the microfile with the greatest negative valence, for strategies No. 3 and No. 4—with the smallest negative one.
2. On step 4, for all strategies we try out all the possible vertices, and choose the one that guarantees the minimum values of (17) on step 6.

4.4 Practical Results of Providing Anonymity for the Fuzzy Group of Military Enlisted Members

To solve the TPGA for the group of military enlisted members (Sect. 3.2) at the first stage, we need to obtain the modified concentration surface according to the procedure described in Sect. 4.2.2. We chose the Daubechies tenth-order wavelet decomposition filters [27] to perform WT. Applying (14) to $\mathbf{C}$ from Sect. 3.2, we obtain the following approximation coefficients of the first decomposition level:

$$\mathbf{a}_1 = \begin{pmatrix} 0.016 & 0.019 & 0.007 & 0.022 & 0.021 & 0.019 & 0.018 & 0.015 & 0.019 \\ 0.029 & 0.033 & 0.018 & 0.037 & 0.035 & 0.031 & 0.031 & 0.030 & 0.030 \end{pmatrix}.$$

As we recall from Sect. 3.2, there are extreme values in the first, second, and fifth columns of $\mathbf{C}$. One of the ways to mask them is to use such modified coefficients (we present them with two decimal points due to space limitations):

$$\tilde{\mathbf{a}}_1 = \begin{pmatrix} 54.72 & -134.57 & 85.97 & 118.03 & 213.19 & -106.42 & -7.61 & 42.90 & 253.79 \\ -7.71 & 113.88 & 227.45 & -83.60 & -15.03 & 28.54 & 280.46 & 28.82 & -106.33 \end{pmatrix}.$$

Applying the generalized version of (10), we obtain the new surface approximation $\tilde{\mathbf{A}}_1$. By adding this approximation to the old surface details $\mathbf{D}_{h1}$, $\mathbf{D}_{v1}$, and $\mathbf{D}_{d1}$ according to (15), we obtain the surface $\breve{\mathbf{C}}$. Since this surface contains negative values, we apply to it (16) ($\gamma = \frac{\sum_{i=1}^{4}\sum_{j=1}^{18} c_{ij}}{(4\times 18)} - \min\left(\breve{\mathbf{C}}\right)$), and obtain the modified concentration surface $\mathbf{C}^*$

$$(\mathbf{C}^{*})^{T} = \begin{pmatrix} 0.001 & 0.008 & 0.017 & 0.010 \\ 0.004 & 0.010 & 0.019 & 0.012 \\ 0.011 & 0.014 & 0.018 & 0.015 \\ 0.013 & 0.011 & 0.008 & 0.011 \\ 0.015 & 0.008 & 0.000 & 0.007 \\ 0.021 & 0.013 & 0.003 & 0.011 \\ 0.024 & 0.018 & 0.011 & 0.017 \\ 0.015 & 0.014 & 0.013 & 0.014 \\ 0.004 & 0.009 & 0.014 & 0.010 \\ 0.001 & 0.010 & 0.021 & 0.012 \\ 0.004 & 0.013 & 0.024 & 0.015 \\ 0.008 & 0.012 & 0.016 & 0.012 \\ 0.012 & 0.009 & 0.006 & 0.009 \\ 0.017 & 0.010 & 0.001 & 0.008 \\ 0.019 & 0.012 & 0.004 & 0.011 \\ 0.020 & 0.015 & 0.009 & 0.014 \\ 0.017 & 0.016 & 0.016 & 0.016 \\ 0.008 & 0.013 & 0.018 & 0.014 \end{pmatrix}.$$

Its details are equal to the details of the initial surface **C** multiplied by the factor of $11,736.620$, i.e. are modified proportionally, which totally suits our purposes of preserving data utility.

Using the inverse of (4) with (5), we obtain the surface $\check{\mathbf{Q}}$, applying (16) (with $\gamma = 0$) with the subsequent rounding to which yields the modified surface $\mathbf{Q}^{*}$:

$$(\mathbf{Q}^{*})^{T} = \begin{pmatrix} 7 & 64 & 134 & 77 \\ 35 & 103 & 186 & 118 \\ 98 & 125 & 159 & 131 \\ 136 & 113 & 85 & 108 \\ 358 & 197 & 0 & 161 \\ 198 & 122 & 28 & 104 \\ 152 & 115 & 70 & 107 \\ 714 & 673 & 622 & 664 \\ 143 & 317 & 531 & 356 \\ 7 & 93 & 199 & 113 \\ 34 & 112 & 207 & 129 \\ 64 & 97 & 137 & 104 \\ 100 & 75 & 46 & 70 \\ 177 & 105 & 16 & 88 \\ 195 & 125 & 40 & 110 \\ 439 & 337 & 212 & 314 \\ 479 & 466 & 450 & 463 \\ 351 & 541 & 772 & 584 \end{pmatrix}.$$

The sum of rows of $\mathbf{C}^{*}$ is shown in Fig. 8 along with the superimposed quantity signal **q** from Sect. 3.1 (to fit the scale, we once again normalized each of two vectors

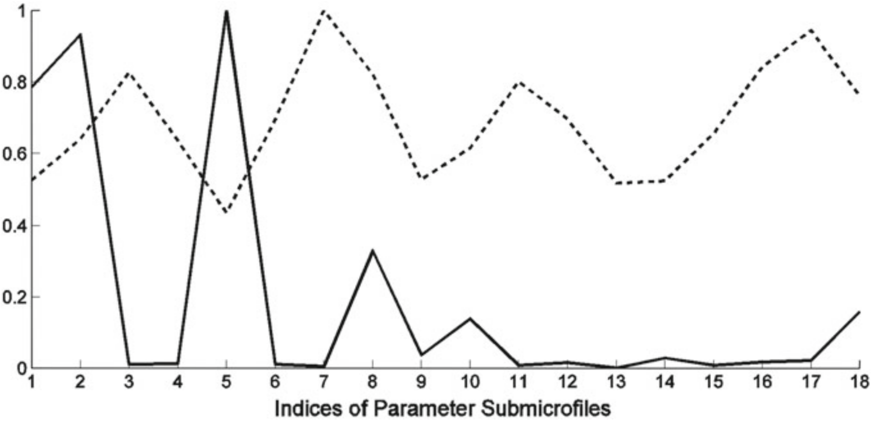

Fig. 8 The initial quantity signal (*solid line*) and the sum of the rows of the modified concentration surface (*dashed line*) for the example

Table 5 Results of applying heuristic strategies to modifying the microfile

Strategy number	Cost of full and partial swapping plans	Cost of transferring plan
1	1,931	14,466
2	2,112	14,306
3	1,931	14,535
4	2,116	14,347

by dividing them by their maximal values). As we can see, extreme values in the first, second, and fifth signal elements have been successfully masked.

Now we need to modify the microfile in order to adjust it to the modified quantity surface $\mathbf{Q}^*$. To perform microfile modification according to Sect. 4.3.2, we took microfile attributes “Sex,” “Age,” “Black of African American,” “Marital Status,” “Educational Attainment,” “Citizenship Status,” “Person’s Total Income in 1999,” and “Hours per Week in 1999” as the influential ones. For the sake of simplicity, we considered every attribute to be nominal, and we assumed $\gamma_k = 1 \; \forall k = \overline{1,8}$, $\chi_1 = 1$, $\chi_2 = 0$. In this case, (17) shows the overall number of attribute values to be changed in order to provide group anonymity.

The results of applying strategies No. 1–4 to the modified quantity surface are presented in Table 5. Since there are 278,337 respondents that have a positive grade of membership in the fuzzy group of the military enlisted members, and we took 8 influential attributes, we see that to provide anonymity we need to alter at most only $\frac{(1931+14535)}{(8\times 278337)} = 0.007$ of all microfile attribute values.

5 Conclusion and Future Research

In the chapter, we showed that microfiles could be considered an important source of information during cyber warfare. We proposed a generic approach to violating anonymity of crisp and fuzzy groups of respondents, and illustrated the importance of such problems with the real data based example concerning violating anonymity of the fuzzy group of "respondents who can be considered military enlisted members with the high level of confidence." We showed that the group anonymity in this case could be provided by modifying values of about 0.7 % of all the microfile attribute values, which is an acceptable cost in most practical situations.

We believe the research can be continued in the direction of developing efficient algorithms for the second stage of solving the TPGA, including evolutionary computation methods. In addition, it is important to enhance the proposed method for constructing FIS for defining fuzzy respondent groups by applying neural network technologies for defining parameters of membership functions.

References

1. Gantz, J., Reinsel, D.: Big data, bigger digital shadows, and biggest growth in the Far East. http://www.emc.com/leadership/digital-universe/iview/executive-summary-a-universe-of.htm (2012)
2. Pfitzmann, A., Hansen, M.: A terminology for talking about privacy by data minimization: anonymity, unlinkability, undetectability, unobservability, pseudonymity, and identity management, Version v0.34, http://dud.inf.tu-dresden.de/Anon_Terminology.shtml (2010)
3. Chertov, O., Tavrov, D.: Data group anonymity: general approach. Int. J. Comput. Sci. Inf. Secur. **8**(7), 1–8 (2010)
4. Chertov, O. (ed.): Group Methods of Data Processing. Lulu.com, Raleigh (2010)
5. Sweeney, L.: Computational Disclosure Control: A Primer on Data Privacy. Ph.D. thesis, Massachusetts Institute of Technology, Cambridge (2001)
6. Evfimievski, A.: Randomization in privacy preserving data mining. ACM SIGKDD Explor. Newslett. **4**(2), 43–48 (2002)
7. Domingo-Ferrer, J., Mateo-Sanz, J.M.: Practical data-oriented microaggregation for statistical disclosure control. IEEE Trans. Knowl. Data Eng. **14**(1), 189–201 (2002)
8. Fienberg, S.E., McIntyre, J.: Data swapping: variations on a theme by Dalenius and Reiss. In: Domingo-Ferrer, J., Torra, V. (eds.) Privacy in Statistical Databases, PSD 2004. LNCS, vol. 3050, pp. 14–29. Springer, Berlin (2004)
9. Wang, J., Zhong, W., Zhang, J.: NNMF-based factorization techniques for high-accuracy privacy protection on non-negative-valued datasets. The 6th IEEE International Conference on Data Mining Workshops. ICDM Workshops 2006, Hong Kong, December 2006, pp. 513–517. IEEE Computer Society Press, Washington (2006)
10. Xu, S., Zhang, J., Han, D., Wang, J.: Singular value decomposition based data distortion strategy for privacy protection. Knowl. Inf. Syst. **10**(3), 383–397 (2006)
11. Liu, L., Wang, J., Zhang, J.: Wavelet-based data perturbation for simultaneous privacy-preserving and statistics-preserving. In: 2008 IEEE International Conference on Data Mining Workshops, Pisa, December 2008, pp. 27–35. IEEE Computer Society Press (2008)
12. National Institute of Statistics and Economic Studies. Minnesota Population Center. Integrated Public Use Microdata Series, International: Version 6.2 [Machine-readable database]. University of Minnesota, Minneapolis, https://international.ipums.org/international/ (2013)

13. Nuclear Power in France, World Nuclear Association, http://www.world-nuclear.org/info/inf40.html
14. Chertov, O., Tavrov, D.: Group anonymity. In: Hllermeier, E., Kruse, R., Hoffmann, F. (eds.) Information Processing and Management of Uncertainty in Knowledge-Based Systems. Applications. CCIS, vol. 81, pp. 592–601. Springer, Berlin (2010)
15. Chertov, O., Tavrov, D.: Group anonymity: problems and solutions. Lviv Polytechnic Natl. Univ. J. Info. Syst. Netw. **673**, 3–15 (2010)
16. Chertov, O., Tavrov, D.: Providing data group anonymity using concentration differences. Mathe. Mach. Syst. **3**, 34–44 (2010)
17. Tishchenko, V., Mladientsev, M.: Dmitrii Ivanovich Miendielieiev, yego zhizn i dieiatielnost. Univiersitietskii pieriod 1861–1890 gg. Nauka, Moskva (1993) (In Russian)
18. Zadeh, L.A.: The concept of a linguistic variable and its application to approximate reasoning. Inf. Sci. **8**, 199–249 (1975)
19. Chertov, O., Tavrov, D.: Providing Group Anonymity Using Wavelet Transform. In: MacKinnon, L.M. (ed.) Data Security and Security Data. LNCS, vol. 6121, pp. 25–36. Springer, Berlin (2012)
20. Mamdani, E.H., Assilian, S.: An experiment in linguistic synthesis with a fuzzy logic controller. Int. J. Man-Mach. Stud. **7**(1), 1–13 (1975)
21. Klir, G.J., Yuan, B.: Fuzzy Sets and Fuzzy Logic: Theory and Applications. Prentice Hall, Upper Saddle River (1995)
22. Zadeh, L.A.: Outline of a new approach to the analysis of complex systems and decision processes. IEEE Trans. Syst. Man Cybern. SMC-**3**(1), 28–44 (1973)
23. U. S. Census 2000. 5-Percent Public Use Microdata Sample Files, http://www.census.gov/main/www/cen2000.html
24. Demographics. Profile of the Military Community. Office of the Deputy under Secretary of Defense (Military Community and Family Policy), http://www.militaryonesource.mil/12038/MOS/Reports/2011_Demographics_Report.pdf (2012)
25. Mallat, S.: A Wavelet Tour of Signal Processing. Academic Press, New York (1999)
26. Chertov, O.R.: Minimizatsiia spotvoren pry formuvanni mikrofailu z zamaskovanymy danymy. Visnyk Skhid-noukrainskoho Natsionalnoho Universytetu imeni Volodymyra Dalia, **8**(179), 256–262 (2012) (In Ukrainian)
27. Daubechies, I.: Ten lectures on wavelets. Soc. Ind. Appl. Math. (1992)