UDC 004

**O. Chertov**
**D. Tavrov**

# DATA GROUP ANONYMITY IN MICROFILES

National Technical University of Ukraine "Kyiv Polytechnic Institute"

*In the paper, we consider practical aspects of providing data group anonymity. We propose a formal definition of the task, discuss its main peculiarities and types, and examine a problem of preventing data utility loss. In particular, we address the problem of minimizing data distortion when perturbing microfile data. For that matter, we introduce a special metric and propose a strategy for its applying to producing optimal microdata distortion.*

***Keywords:*** *group anonymity, statistical disclosure control, privacy protection, microfile, data utility.*

## Introduction

According to the famous Abraham Lincoln's saying, "You can fool some of the people all the time, and all of the people some of the time, but you cannot fool all of the people all the time". In terms of data anonymity, this should sound as "You can hide some of the data all the time, and all the data some of the time, but you cannot hide all the data all the time". But, although one cannot reach total anonymity, there are many cases when one needs to hide (or mask) specific private information on a particular respondent, his intensions and activities. Thus, the legal system of every state includes requirements to provide published data depersonalization and anonymity (e.g., see the Health Insurance Portability and Accountability Act of 1996 (HIPAA) [1], and the Patient Safety and Quality Improvement Act of 2005 (PSQIA) [2] concerning health protection data in the US, or Directive on privacy and electronic communications [3] about electronic commerce in the EU, or the State Statistics Law [4] about providing confidentiality of the primary statistical information in Ukraine).

But, what is usually called data anonymity [5] is in fact individual data anonymity. Individual anonymity methods are used to disable identifying information on a particular respondent in the given data set (e.g., in a microfile) which contains primary and non-aggregated respondent data.

In our previous work [6], we introduced a totally different kind of anonymity, namely, *group anonymity*. It lies in protecting important data features, distributions, and collective patterns which cannot be defined by analyzing individual data records only.

For instance, given a regional distribution of French power plant workers (see the microfile [7] with the 1999 population census data), we can almost ideally determine where French nuclear reactors are located. This is mainly because the nuclear power is a main source of electricity in France. Figure 1 presents a corresponding regional distribution of the power plant workers (the quantities above the solid line can reveal the nuclear power plants locations).

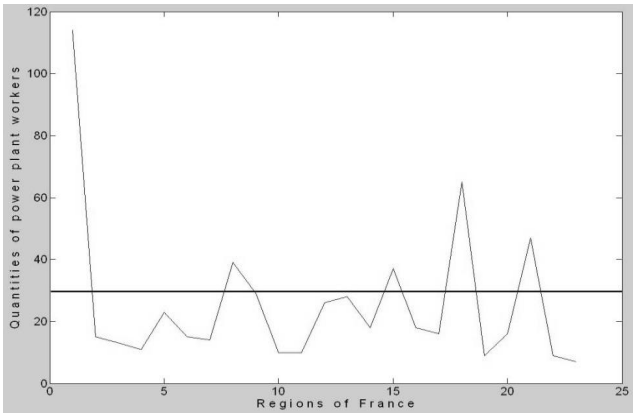


Fig. 1. Power plant workers and power plants distribution over the regions of France.

Now, let's suppose the French government decides to build a secret nuclear research center in a region with no nuclear plants registered officially. If the microfile with the census data isn't protected, we might easily track this center using Fig. 1.

Surely, this example is a bit far-fetched. Nevertheless, the confidential information of such a kind can also be revealed using various indirect data. For instance, the site of a military base can be determined if the density of young males is higher than the density of young females. Or, a relatively big number of respondents with a degree in science working in a region without any prominent educational or research centers might indicate a hidden scientific lab.

Thus, the task of providing anonymity for specific (*vital*) value combinations distribution over determined (*parameter*) values is important and topical. This is especially the case when preparing microfiles containing various demographic, economical and other types of information. In this paper, we address the practical aspects of providing such data group anonymity.

When publishing primary data, it is necessary to provide anonymity taking into consideration the adversarial model. It reflects which additional information might be known to a potential attacker for violating privacy in the published data. In this paper, we will suppose that the potential adversary does not possess any additional data except for those ones contained in the masked microfile.

The rest of the paper goes as follows. In Section 2, we provide a general survey on the published works in this field. Also, we discuss the existing terminology. Section 3 provides necessary definitions, a formal problem set, and a bunch of ways to solve it. Different kinds of group anonymity task are discussed in Section 4. Section 5 is dedicated to the microfile utility preservation. In Section 6, we address the problem of modifying the microfile with a minimal distortion. In the last section, we draw conclusions and set the tasks to be solved in the future.

**Related Work**

**1. Anonymity**

Anonymity is derived from the Greek word бнщнхмйб meaning "without a name" or "namelessness". In a consolidated proposal for terminology [5], anonymity of a subject means that the subject is not identifiable (uniquely characterized) within a set of subjects. Then, data anonymity, as we understand it in this paper, means that the respondents' records cannot be identified within a microfile.

Basing on the model of an anonymous message transmission system, there can be distinguished two anonymity classes, namely, global and individual [5] (or local [8]) anonymity. Whether the anonymity belongs to a particular class depends on its purpose. It can be achieved for the users corresponding to the message (senders or recipients) globally (for all such users), or for each user individually. In terms of the task under review in this paper, such classes won't be differentiated [9], and will be considered as the anonymity of individuals or *individual data anonymity* (property of being unidentifiable within a group).

We have to pay attention to several other important positions connected with the term of anonymity. When the systems for anonymous communication were studied, the term of the degree of anonymity was introduced [10]. If to apply this term to the field under review in this paper, we can admit that the degree of anonymity provided against an attacker can be viewed as a continuum, ranging from complete anonymity ("absolute privacy") (when the attacker cannot perceive the presence of some respondents in the data sample) to no anonymity ("probably exposed") (when the attacker can identify the personal respondent record).

While discussing the term of anonymity, it is important to note that there exists a set of similar (and sometimes even overlapping) but rather different information-hiding requirements [11], [12], [5] such as noninterference, privacy, confidentiality, secrecy, unlinkability etc., but they are not the subject to this paper.

**2. Individual and group data anonymity**

The restrictions required to provide the data anonymity of individuals are important and comprehensive. That is why, they are having been successfully studied for many years now by the researchers in such fields as statistical disclosure control [13], privacy-preserving data mining [14], distributed privacy, cryptography and adversarial collaboration [15].

Nowadays, the data anonymity is achieved by using perturbative and/or non-perturbative methods for masking microfile data (as it can be found in [16]). The perturbative methods modify particular microfile values (for example, consider data swapping [17], microaggregation [18], rounding [16] and others). At the same time, such methods as generalization, top and bottom coding, and local suppression described in [13] belong to the non-perturbative ones as they let to mask some value combinations without altering them.

There have been also determined two main principles for providing individual data anonymity, i.e. randomization and *k*-anonymity. The main idea of the data randomization (refer to [14], [19], [20], or rather close works on differential privacy [21]) is to add noise to the data so that records' attribute values are masked. Another principle discussed in [9], [22] (namely, *k*-anonymity) means that every attribute values' combination corresponds to at least *k* respondents with the same combination in the microfile.

Some authors [23] call the situation when it is impossible to find out a specific individual among a respondent group as group anonymity. But, in our opinion, this name is not suitable for this situation because it concerns the individual respondent anonymity inside a specific group rather than anonymity of a respondent group itself.

In the past few years, methods and tools for publishing useful information while preserving data

privacy are usually viewed as privacy-preserving data publishing. The work [24] contains a systematized survey of the researches in this field.

Recently, two novel techniques have been proposed to provide data anonymity. The first one is based on the structure transformations of the data matrix (it could possibly be the microfile data). Two methods representing such a technique are singular value decomposition [25] and nonnegative matrix factorization [26]. The second technique implies using Fourier transformation [27] or wavelet transformation [28], [29].

But, all the discussed principles and techniques intend to provide mainly the data anonymity of individuals. Although, they are completely useless (at least in their today form) for providing group anonymity. This task was initially introduced in [30]. In [6], this term was specified. Also, there was proposed a method for providing data group anonymity using wavelet transform.

**Group Anonimity Definition and Ways to Provide It**

A microfile can be viewed as a matrix **M**, with each row unambiguously corresponding to one record. Every row itself can unambiguously be matched with only one respondent.

We will treat each **M** column as a set of values corresponding to some respondent attribute. For a person, these attributes could possibly be "age", "name", "sex" and others.

Usually [31], there are defined two basic groups of attributes.

*An identifier* is an attribute which unambiguously defines (or identifies, which follows from its name) a certain respondent in a microfile (e.g., "Social Security Number" or "Full Name"). In fact, the only way to prevent privacy violation in this case is to completely remove the identifiers.

Another set of attributes called *quasi-identifiers* consists of those ones which can be joined with external information to re-identify respondents (with a sufficiently high probability) [9].

We will additionally name all the other attributes *ancillary* ones. For instance, it might probably be "language skills". These attributes do not play any prominent role when providing group anonymity, but they have to be considered to guarantee minimal microfile distortion.

Still, this division is appropriate for individual anonymity. But, it's completely insufficient when hiding, say, regional distribution of some respondents. That's why we will distinguish among the quasi-identifiers and ancillary attributes those ones whose distribution has to be hidden (masked) in terms of group anonymity. We will call such attributes and corresponding value combinations the *vital* ones. The values (and appropriate attributes) to distribute over will be called the *parameter* ones. For example, when masking regional distribution of military officers the vital attribute would be "Military Status", whereas "Place of Work" would be a parameter one.

We will call *a goal signal* a numeric vector $\theta = (\theta_1, \theta_2, ..., \theta_n)$. It consists of the quantities of microfile records with a specific vital value combination and a parameter value.

Additionally, we will call *a goal chart* a graphical representation of a goal signal (for instance, see Fig.1).

Depending on a parameter attribute, this chart can be either the cross-sectional data (like a distribution of companies by their product yield) or the time-series (like respondents' answers to the same questions during a long-running opinion poll). In some way, these charts resemble the ones considered in technical analysis [32].

Every chart (distribution) is defined by a list of its main features. Their importance for providing group anonymity is presented in Table 1.

Having introduced all the necessary terms, we can provide a formal data group anonymity definition.

*Definition.* Let us be given a microfile **M** with all the identifiers having been eliminated. Among the remaining attributes $\{A_1, A_2, ..., A_l\}$ we will define groups $\{G_i\}$, $i = 1,...,k$. Each of them consists of one or several vital attributes $\{V_1^{(i)}, V_2^{(i)}, ..., V_m^{(i)}\}$ and one parameter attribute $P^{(i)}$, $P^{(i)} \neq V_j^{(i)}$, $j = 1,...,m$. Then, the task of *providing data group anonymity* lies in the following. For each group $\{G_i\}$, $i = 1,...,k$, we need to modify a goal signal such way that at least one main feature of a goal chart changes so that it leads to a *needed* distribution masking. At the same time, we have to preserve the goal signal's mean value:

$$\sum_i \theta_i = \sum_i \theta_i^* . \quad (1)$$

In (1) and further on, an asterisk * marks the modified data elements.

Table 1

Main goal chart features from a group anonymity point of view

| № | Main Chart Features | Importance for Group Anonymity |
|---|---|---|
| 1 | extremums | almost always this is the most confidential feature |
| 2 | statistical features | this can be topical when a parameter attribute is numerical or ordinal; useless when they are nominal |
| 3 | trends | this can be important if the parameter values are numerical or ordinal, and the trend is somewhat confidential |
| 4 | cycles or periods (both total and partial) | this is often important when a parameter attribute is temporal |
| 5 | frequency spectrum | this can be necessary if the distribution contains the parts repeated cyclically |

Let us consider one example. Suppose we set a task of protecting the regional distribution of scientific professionals and technicians in the UK. The importance of the task is clear since maximums in an appropriate goal signal could possible lead to revealing the locations of restricted scientific centers. We can take "Region of the UK" as a parameter attribute. Also, we can take "Occupation" as a vital one. But, since we are concerned with redistributing people of a particular occupation, we choose only two vital values standing for "Science Professionals" and "Science and engineering Technicians".

Depending on the goal chart features to be masked, the data group anonymity task can be solved in different ways (see Table 2).

Table 2

Different ways to provide group anonymity

| № | Main Chart Features | Some Ways to Provide Group Anonymity |
|---|---|---|
| 1 | extremums | transiting extremums;putting several alleged ones |
| 2 | statistical features | altering them to receive another statistical signal distribution |
| 3 | trends | changing inclination angle for a linear one; changing the degree of a trend polynomial |
| 4 | cycles or periods (both total and partial) | changing a period of a signal (or its part); changing an amplitude of the repeated signal part; adding new periodic distribution components |
| 5 | frequency spectrum | changing a period of a signal (or its part); adding new periodic distribution components; changing a wavelet approximation of a signal |

Of course, in practice, we might use more than one of the ways listed above. Besides, one way can sometimes imply changing some signal features. E.g., changing a wavelet approximation can also lead to extremum transitions [6] or creating some new alleged extremums.

It seems unpromising to try to find any common group anonymity measure for all the methods listed in Table 2. Possibly, the solution should be only qualitative.

**Different Kinds of Group Anonymity Task**

According to Section 1, the confidential information sometimes can be revealed without analyzing the direct data. E.g., appropriate relative ratios or information on exceeding a certain threshold can pose a much bigger privacy violation threat than the absolute quantities.

In such cases, the goal signal has to contain somehow transformed absolute quantities of records with specific vital value combinations.

The easiest form of a goal signal is obviously a *quantity* signal consisting of the absolute quantities discussed above:

$$q = (q_1, q_2, ..., q_n). \quad (2)$$

This type of a goal signal is most sufficient in situations when the very presence of the records in a data set is confidential. For instance, this is the case with protecting the regional distribution of the military personnel, or the distribution of people with HIV over different ethnic groups.

In different other cases the confidential information can be contained not in the absolute data but in the relative ones. To protect such information, we need to build up a *concentration* signal:

$$c = (c_1, c_2, ..., c_n) \equiv \left( \frac{q_1}{\rho_1}, \frac{q_2}{\rho_2}, ..., \frac{q_n}{\rho_n} \right). \quad (3)$$

In (3), $\rho_i$ stands for a number of records with the vital attribute's values which comprise the vital ones. For example, if a vital value combination is (female; 30 years old) then the values $\rho_i$ might

stand for women of any age, or for all the people in general.

Also, the third typical situation arises when the confidential information can be determined through the comparative data. In this case, we need to define two different vital value combinations and build two different concentration signals, $c^{(a)} = \left(c_1^{(a)}, c_2^{(a)}, ..., c_n^{(a)}\right)$ and $c^{(b)} = \left(c_1^{(b)}, c_2^{(b)}, ..., c_n^{(b)}\right)$. Then, the goal signal can be defined as a *concentration difference* signal:

$$\delta \quad (\delta_1, \delta_2, ..., \delta_n) \equiv \left(c_1^{(b)} - c_1^{(a)}, c_2^{(b)} - c_2^{(a)}, ..., c_n^{(b)} - c_n^{(a)}\right). \tag{4}$$

This kind of a goal signal is useful to protect the site of a military base which can be obtained by analyzing the differences between young males' and females' concentration in every region.

**Preserving Masked Microfile Utility**

According to Table 2, group anonymity can be achieved by changing different main goal chart features. On the other hand, it is obvious that it is better to change a chart feature, and preserve some other ones at the same time.

Actually, there have already been developed several techniques to provide group anonymity which guarantee preserving some goal chart features.

One of them lies in changing chart's extremums with simultaneous preservation of its main statistical features as mean value and standard deviation. This can be done by *normalizing* the modified goal signal according to the following formula [29]:

$$\theta^* = \left(\theta + \frac{\sigma^*}{\sigma} \cdot \mu - \mu^*\right) \cdot \frac{\sigma}{\sigma^*}. \tag{5}$$

Here, $\mu = \frac{1}{n}\sum_{i=1}^{n}\theta_i$, $\mu^* = \frac{1}{n}\sum_{i=1}^{n}\theta^*_i$, $\sigma = \sqrt{\frac{\sum_{i=1}^{n}(\theta_i - \mu)^2}{n-1}}$, $\sigma^* = \sqrt{\frac{\sum_{i=1}^{n}(\theta^*_i - \mu^*)^2}{n-1}}$.

Another way to provide group anonymity lies in changing chart's extremums with simultaneous preservation of its frequency features. This approach is thoroughly discussed in [6], and is based on using wavelet transform:

$$\theta(t) = \sum_i a_{k,i} \cdot \varphi_{k,i}(t) + \sum_{j=k}^{1}\sum_i d_{j,i} \cdot \psi_{j,i}(t). \tag{6}$$

Here, $\varphi_{k,i}$ stands for a shifted and sampled scaling function, whereas $\psi_{j,i}$ stands for a shifted and sampled wavelet function.

It was shown [6] that to change approximations we need to modify approximation coefficients $a_{k,i}$. At the same time, leaving detail coefficients $d_{j,i}$ permanent preserves important signal's frequency features.

Of course, there exist other approaches apart from (5) and (6) that provide group anonymity and leave main chart's features unchanged.

**Minimal Distortion as a Criterion for Choosing Particular Microfile Perturbing Techniques**

When solving a group anonymity task, we are obliged to distort the primary data. In general, the more distortion is introduced, the highest privacy protection quality is achieved. But, the data utility loss might also become greater.

For us, the primary data is modified the best way if the distortion in the microfile is as little as possible. Thus, we set a following optimization task. We need to provide group anonymity by making *as few changes* to a microfile as can be.

Actually, the aim of modifying the microfile is to adapt it to the modified goal signal. This procedure can be treated as changing parameter values of a particular number of records necessary for obtaining the needed distribution defined by the goal signal.

Also, such parameter value modifications should be carried out simultaneously for two different records, so that the number of records with a particular parameter value persists. In general, we can choose any pair of records for such an operation (as long as only one of them contains a vital value combination). But, in this case obtaining new distributions of vital values can spoil the distribution of the other ones.

We will call the attributes whose distribution is of a big importance for a data recipient *influential attributes* (the vital attributes are influential by definition). Then, under minimal microfile distortion we will

understand swapping the parameter values between two microfile records which are close to each other.

To define how close the records are, we introduce an *influential metric*:

$$InfM(r,r^*)=\sum_{p=1}^{n_{ord}}\omega_p\left(\frac{r(I_p)-r^*(I_p)}{r(I_p)+r^*(I_p)}\right)^2+\sum_{k=1}^{n_{nom}}\gamma_k\left(\chi(r(J_k),r^*(J_k))\right)^2 \quad (7)$$

Here, $I_p$ stands for the $p^{\text{th}}$ ordinal influential attribute (making a total of $n_{ord}$), $J_k$ stands for the $k^{\text{th}}$ nominal influential attribute (making a total of $n_{nom}$). Also, $r(\bullet)$ stands for an appropriate attribute value of a record $r$. The operator $\chi(v_1,v_2)$ equals $\chi_1$ if two values $v_1$ and $v_2$ fall into one category, and $\chi_2$ in the other case. The weights $\omega_p$ and $\gamma_k$ regulate the importance of each attribute (e.g., for the attributes not to be changed they should be maximal).

Having defined metric (7), we can propose a strategy for constructing a new microfile which lies in examining all potential record pairs $\langle r,r^*\rangle$ to swap their parameter values and choosing those ones with the lowest $InfM(r,r^*)$ value.

Since such a strategy seems to be NP-hard, there can also be developed different heuristics providing not optimal but quite acceptable solutions. Though, this is a subject to another discussion.

The mentioned optimization task can be solved either for a standalone group $G_i$, or simultaneously for several ones. But, in this case the task can be replaced with the one of not exceeding certain fixed thresholds, and in this situation minimal distortion of each group can be calculated using different weights for different groups.

**Conclusions and Future Research**

Providing both individual and group anonymity before publishing the microfile data is a task of a great importance. But, despite the fact that there have been constructed different methods and systems for protecting individuals, the problem of providing data group anonymity is only at its primary stage of development.

In this paper, it is the first time when the term of data group anonymity has been formally introduced. Also, we've classified different ways of its achieving through modifying the distributions of value combinations to be protected. We called them vital value combinations. At the same time, we especially drew attention to the practical aspects of implementing data group anonymity in microfiles. In particular, we've introduced an influential metric which can be used to set an optimization task of minimizing the primary microfile data distortion.

We think that it is necessary to continue studying the group anonymity in two directions:

- finding other, different from mentioned in Section 5, ways of preserving perturbed microfile utility;
- developing heuristic methods to provide acceptable data distortion while masking the microfile.

**References**

1. Health Insurance Portability and Accountability Act of 1996 (HIPAA): Aug. 21, 1996 / 104th Congress. – Public Law 104-191. – Режим доступу: http://www.hipaa.org/.

2. Patient Safety and Quality Improvement Act of 2005 (PSQIA) / Federal Register. – 2001 – 73(266).

3. Directive 2002/58/EC of the European Parliament and of the Council of 12 July 2002: Jul. 31, 2002 / Official Journal of the European Communities. – 2002 – L 201.

4. Закон України "Про державну статистику": станом на 5 бер. 2009. – Режим доступу: http://zakon1.rada.gov.ua/cgi-bin/laws/main.cgi?nreg=2614-12\&p=1265575855780241.

5. A Terminology for Talking about Privacy by Data Minimization: Anonymity, Unlinkability, Undetectability, Unobservability, Pseudonymity, and Identity Management, Version v0.32 [Електронний ресурс] / A. Pfitzmann, M. Hansen. – 2009. – Режим доступу: http://dud.inf.tu-dresden.de/Anon_Terminology.shtml.

6. Chertov, O., Tavrov, D. Group Anonymity // Ньllermeier, E., Kruse, R. (eds.) IPMU-2010. – Heidelberg: Springer, 2010. – CCSI, vol. 81. – P. 592-601.

7. Minnesota Population Center. Integrated Public Use Microdata Series International [Електронний ресурс]. – Режим доступу: https://international.ipums.org/international/