The Anonymization of Files with Itemized Information

Ariel Mansura*

Abstract

The Bank of Israel's Information and Statistics Department gathers and manages files from a variety of sources, some of which contain itemized information. In order to enable freedom of information while also maintaining the confidentiality of the information, the Department builds anonymization procedures for the information files. This is a complex process, the aim of which is to prevent the identification or disclosure of sensitive or confidential information on individuals whose data appear in the files. This work outlines the process of anonymization of itemized data, defines the basic terms involved, presents accepted methods for assessing the risk of disclosure in the files, and samples the implementation of the process.

* Bank of Israel Information and Statistics Department.

1. INTRODUCTION

Following the Global Financial Crisis of 2008, central banks, including the Bank of Israel, began managing macroprudential policy, the aim of which is to identify systemic risks at the formative stage and to advance actions that will deal with them and limit their effect on the financial stability of the economy. The new challenges are motivating the central banks to manage consistent and integrative databases that will support this policy. Alongside technological development, which makes it possible to store and process very large quantities of information, there is an increasing need for databases of itemized data, which will enable the completion of information on the flow of capital in the economy, and on which bases it will be possible to obtain a detailed and available picture of the state of financial stability and robustness.

Against the background of these trends, and in parallel with the development of freedom of information laws that emphasize the importance of increased transparency and sharing of information, various entities that manage statistical information tend to enable access to itemized information as well, for the purposes of managing policy, economic analysis, and research. In order to allow access to such information within the organization or outside it, the Protection of Privacy Law requires that the confidentiality of the information be maintained, as the information relates to individual persons. In addition, the law requires that the commercial confidentiality of business entities be maintained—a complex task, particularly when dealing with financial information that is sometimes characterized by high concentration.

The Information and Statistics Department at the Bank of Israel, which collects and creates financial statistics, manages databases that include, among other things, itemized information on various topics: the capital market, the foreign exchange market, banking, the credit market, and more. In this context, the Bank of Israel is currently building a credit register that includes itemized information on the credit history of borrowers in the economy, and which will help the credit bureaus[1] in building models for the credit rating of borrowers. Based on this register, the Information and Statistics Department will manage a statistical database where the itemized information contained in it is not identified, for the Bank of Israel's internal uses in order to fulfill its legally mandated functions.

In order to enable access to the information, while also maintaining its confidentiality, the Bank of Israel is designing a process called "the anonymization of data files". The objective of the anonymization process is to protect the information so that it will not be possible to identify or expose the individuals whose data appear in the files, particularly information about them that is sensitive or confidential.[2] This process will relate to both data intended for use within the Bank—even though only a few economists within the Bank will be permitted to access them—and information that is permitted to be accessible to researchers from outside the Bank, subject to the privacy protection restrictions and maintaining commercial confidentiality.

A database containing itemized information naturally includes information that directly identifies the individual—a field that on its own exposes the identity of the individual even without needing additional information located in other fields. Examples of this include the identification number and full name of the individual. Therefore, a necessary condition for anonymizing the database is the deletion of all direct identifiers. However, this condition is not sufficient to protect the database, because even without this

[1] The Credit Data Law, Section 16. This law will soon come into force.

[2] The anonymization process described in this work does not relate to series that the Information and Statistics Department publishes on the Bank of Israel's website. Those series present aggregate, and not itemized, information.

information, it is sometimes possible to discover information on individuals by connecting a number of fields, or cross-referencing them with information from other databases the access to which is permitted. An individual can also be identified by searching for combinations that are not common among the relevant population, which are characteristic only of a particular individual or a small group of individuals.

The anonymization process begins with a precise definition of disclosure scenarios (see terms in Section 2). These scenarios include the possibilities available to users in order to expose information on individuals, and against which we want to be protected. With the given scenarios, we can use methods to blur the identification and protect the information. At the end of the process, we will have to assess the remaining risk and quantify the information that was lost as a result of the process. It is clear that there is a tradeoff between the extent of anonymization, meaning the extent of protection of the file, and the extent of usability of the data, since there is a loss of information.

This work will describe the anonymization process for itemized data, define the basic terms on the subject, present accepted methods for assessing the risk, and sample the implementation of the process on sample tables of data that include itemized information.[3]

2. TERMS THAT ARE RELEVANT TO THE ANONYMIZATION PROCESS

Statistical disclosure control – A general term describing the group of methods for reducing the risk of disclosure (hereinafter “disclosure”) of individuals in the file. In general, the methods are divided into two: a. Perturbative methods (which add noise to the data); and b. Nonperturbative methods such as grouping categories of fields or deleting values from fields with higher than permitted risk and inserting missing values in their place.

Anonymization – A process in which an unprotected file becomes a protected file according to the protection level set out in advance through the statistical disclosure control.

Disclosure[4] **-** The disclosure of information that was not known and published beforehand regarding an individual through an information file that was distributed. There are three types of disclosure:

- **Identity disclosure** – Connecting the known identity of the individual, such as his name and surname, to a record in the file. When such a connection is made, the information in the other fields in the file of this individual is exposed. For instance: cross-referencing a record with two fields—the individual’s identification name and monthly income—with a record from an external file where the individual’s full name also applies (or with the personal information of the individual with the same identification number) will cause this individual’s identity and monthly income to be exposed.

[3] In order to maintain simplicity and precision as much as possible, we will deal here with data that include all of the records of the relevant population (“census file”) and not a sample file such as a survey that includes only some of the records of the relevant population, where the method of handling is more complex. We will also assume that the file does not have a hierarchical structure, a structure that is characteristic, for instance, of a data file that includes a variable that indicates the household to which the individual belongs.

[4] See, for instance, item [4] in the Bibliography.

Record from the income file	
Identification number	**Monthly income in shekels**
123456	**5,000**

Record from an external file	
Full name	**Identification number**
Yisrael Yisraeli	**123456**

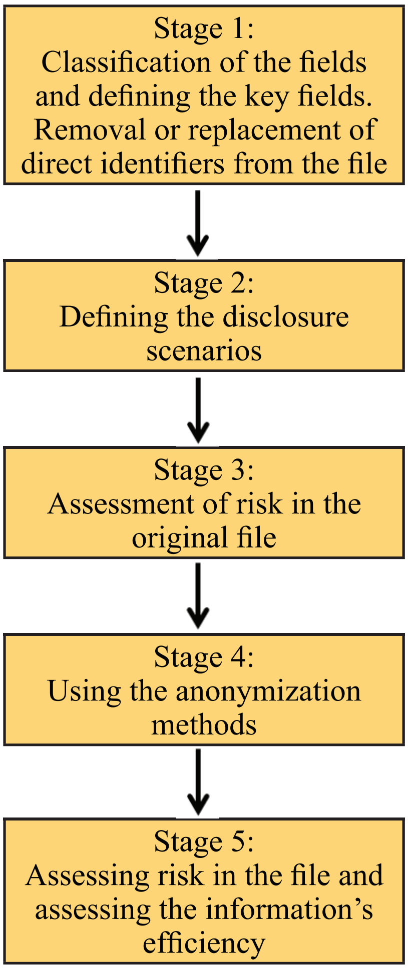
- **Characteristic disclosure** – Disclosing a particular characteristic of the individual even without connecting the individual's identity to a certain record. For instance: if the income of all individuals aged 70–74 in a particular file, with exception, is between 5000 and 10,000 shekels, we can know the income range of an individual whose age is within the age range included in that file, even without knowing his or her identity.

Gender	Age range	Range of monthly income in shekels
Male	**70–74**	**5,000–10,000**
Female	**70–74**	**5,000–10,000**

- **Statistical disclosure** – Identifying the characteristic of a person through a statistical analysis of the file. For instance: a too-precise guess—through a good forecasting model—of the income of a particular person, based on known characteristics of that person that appear in the file.

3. DESCRIPTION OF THE STAGES IN THE ANONYMIZATION PROCESS

Flowchart of the anonymization process


Stage 1:
Classification of the fields and defining the key fields. Removal or replacement of direct identifiers from the file
Stage 2:
Defining the disclosure scenarios
Stage 3:
Assessment of risk in the original file
Stage 4:
Using the anonymization methods
Stage 5:
Assessing risk in the file and assessing the information's efficiency

Stage 1: Classification of the fields and defining the key fields

Types of field—It is common to divide the fields in a file into three types. This division is not necessarily exclusive: A field can belong to more than one type.

- **Direct identifiers**—Fields that identify individuals in the file without using other fields. Examples of such fields are the identification number, the full name, and the precise address. Fields of this type are deleted from the file in the first stage of the anonymization process, or are replaced on a one-to-one basis with other fields that are not identifiers.

- **Key fields**[5]—Fields that can be cross-referenced with external information, such as those in the published or partially published census file, thereby exposing the identity of the individuals behind certain records in the file.
- **Sensitive fields**—Fields where, due to their sensitivity, it is prohibited that their values, regarding each of the individuals whose identity is known in the file, be disclosed. Examples of such fields are a person's state of health or income.

In addition to this division, the fields can be divided into two other types:

- **Categorical fields**—Fields that include a final number (generally a low number) of categories or values. This group can be divided into ordinal fields and nonordinal fields.
- **Continuous fields**—Numerical fields that can be the subject of arithmetical actions. These fields can obtain a large number of values.

Stage 2: Defining disclosure scenarios

Disclosure scenarios[6] are a group of assumptions that describe how a user, or another person exposed to the file, can expose information on individuals from within the file. For instance: A user can cross-reference the information from the file with other information he has through a number of common characteristics, or through information on an individual that he knows and he is aware that this individual is in the file. In that way, he can disclose additional sensitive information about that individual through the characteristics he knows.

The disclosure scenario can for the most part be summed up by determining groups of key fields through which information in the file can be cross-referenced with other external information (a file or personal knowledge), to discover information on individuals through combinations that are characteristic of only a few individuals in the file.

Setting disclosure scenarios is necessary to the anonymization process, since we are trying to protect the information from them. The assessment of the level of risk of information disclosure is also dependent on setting these scenarios, because it is not general, but relates to certain disclosure scenarios. The disclosure scenarios are determined with the help of experts in the relevant content worlds, who know how and through what means a user, or anyone with access to the information, can disclose information on individuals in the file. Even so, even experts in the content worlds do not know all of the information disclosure possibilities, and in certain cases, the tendency is therefore to assume the worst case scenario.

The disclosure scenarios can be less or more severe than the objective information disclosure possibilities, according to the disclosure policy that depends on how the data are used, the purpose of the use, the identity of the users, the severity of the damage inherent in disclosure, and so forth. In this context, it is common to distinguish between scientific use files, which are used by researchers under contract, subject to permissions and restrictions such as working within a physical research room or a virtual research room through remote access, and public use files that have no restriction or control. The policy regarding the information files issued to the public is generally very strict, and requires significant data processing.

[5] See, for instance, item [7] in the Bibliography.
[6] See, for instance, item [7] in the Bibliography.

Stage 3: Assessing the risk of disclosure in the file

As stated, the risk of disclosure relates directly to disclosure scenarios, meaning to groups of key fields (categorical or continuous) that are defined for a certain file. After the key field groups are defined, a number of risk indices can be addressed.

- **The risk of a record in a file**—the probability of matching a certain record in a file to a certain individual whose identity is known. In this context, a distinction should be made between categorical key fields and continuous key fields. In terms of a scenario in which categorical key fields are cross-referenced, there are two common requirements.
- **K-anonymity requirement**[7]—a requirement that in each combination of categorical key fields in groups that are defined in the disclosure scenario, there shall be at least K records with the same combination. In order to check this, a multi-dimensional table (or tables for each disclosure scenario) can be built, in which the number of cells is equal to the number of possible combinations. Based on this table, the disclosure probability of each record can be calculated. The purpose of this requirement is to protect against the disclosure of identity, because if a certain combination from the table relates to only one individual, that combination can be cross-referenced with the same combination in a different table with the same key fields, thereby disclosing the identity of the individual.
- **I-diversity requirement**[8]—another requirement that is meant to protect against disclosure of characteristics. Each cell in the frequency table may have enough records, but regarding a particular sensitive field, there is no variance among those records that belong to the same combination. The I-diversity requirement is that in all possible combinations there should be at least I different values. In a situation where there is no variance, it is enough to know which combination relates to an individual in order to identify that characteristic with certainty, even without knowing that the record relates to him.

The following table presents these two requirements through a simple example of a scenario in which there are only two key fields – gender and age:

Record	Key field 1 – Gender	Key field 2 – Age range	Frequency of the combination in the table	Sensitive field – interest rate (rounded) on the loan	Number of different values
1	Male	50–60	3	2%	2
2	Male	50–60	3	4%	2
3	Male	50–60	3	4%	2
4	Female	40–50	3	2%	1
5	Female	40–50	3	2%	1
6	Female	40–50	3	2%	1

[7] See, for instance, item [7] in the Bibliography.

[8] See item [5] in the Bibliography.

The table shows that the first individual (Record 1) belongs to a cell with the combination: gender=male; age range=50–60. There are three individuals in the table with this combination (Records 1–3). However, for the sensitivity variable, there are two possibilities (interest of 2% or 4%). All of the records in the table fulfill the 3-anonimity requirement, while only Records 1–3 fulfill the 2-diversity requirement.

- **Risk in continuous key fields**—regarding continuous key fields, we cannot build a frequency table, since most of the values appear only once. It is generally customary to assess the risk in these variables based on the extent to which record linkage is possible between the file where we changed the data on continuous variables, such as by adding noise, and the original file.
- **Global risk of each file**—an index that grades the risk level of the entire file, which is calculated on the basis of an aggregation of the disclosure probabilities of the records in the file. An example of such an index is the total disclosure probability in the file, which is equal to the expected value of the identifications in it.

Stage 4: Using the anonymization method

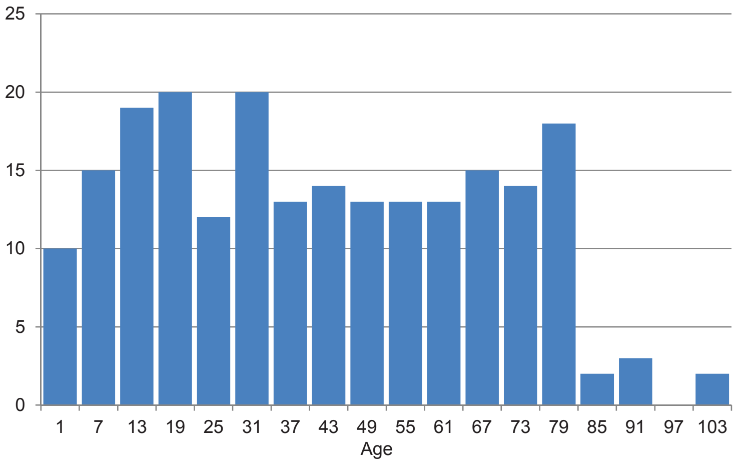
The following is an outline of a number of common anonymization methods:

- **Global recoding**—a method that reduces the level of information in the field and adjusts it to categorical fields and to continuous fields. For a categorical field, global recoding means attaching a number of categories to the common category. For instance, in the field for the individual's profession, the professions of statistician and mathematician can be consolidated into one common category if there are too few records in certain combinations of the key fields that include one of those categories. Another example is changing the age field into ranges of 5 or 10 years. Global recoding in a continuous field is basically replacing a continuous field with a categorical field. For instance, a field that is a loan amount can be replaced by a number of categories that are in ranges of NIS 100,000. The following table shows an example of global recoding of the income field (continuous).

Record number	Monthly income in shekels	Monthly income after recoding
1	8,365	Up to 10,000
2	16,569	10,000–20,000
3	100,200	100,000–200,000
4	5,750	Up to 10,000

- **Upper and lower recoding**—This method is a private case of global recoding, and deals with the tails of the distribution. For a continuous field, it gathers the extreme categories beyond the upper bound of one category, and the same can be done regarding low categories. In a continuous field, the method gathers all the values beyond the upper and/or lower bound of two categories—upper and lower—and in the rest of the range, the data are gathered as in the previous section. This method is appropriate for fields where there are few instances beyond a certain bound. The following figure shows an example of the distribution of the age field, where there are few individuals above age 80. If there are too few records at high ages in the combinations that include the age variable, this method allows us to collect all ages above age 80 into one category—80+.

Figure 1
Distribution of the Age Variable - Sample Data



- **Local suppression**—This method inserts missing values into certain fields of certain records, and is appropriate for categorical fields and not for continuous fields. When there are combinations of key fields where there are few records, a missing value can be inserted in one of the fields. The advantage of this method is that it deals only with records at high risk. On the other hand, it creates a lack of uniformity in a certain field, because a missing value appears in certain records in that field. The following tables shows an example of a local suppression and the insertion of a missing value (NA) in Record 4 regarding the combination of male gender and the 20–30 age range, a combination in which there is only one individual.

	Before local deletion		After local deletion	
Record	**Key field 1 – Gender**	**Key field 2 – Age range**	**Key field 1 – Gender**	**Key field 2 – Age range**
1	**Male**	**50–60**	**Male**	**50–60**
2	**Male**	**50–60**	**Male**	**50–60**
3	**Male**	**50–60**	**Male**	**50–60**
4	**Male**	**20–30**	**Male**	**NA**

- **Adding noise (additive)**[9]—This method changes the numeric values in the field, and is appropriate for continuous fields but not categorical fields. There are a number of accepted paths, two of which are presented below.
- Adding white noise (uncorrelated)—In this method, uncorrelated noise is added to a particular field, which we will label as X, as follows:

$$Z = X + \varepsilon$$

where ε is a vector of normally distributed and uncorrelated noises (white noise). It can be shown that this method maintains (proximately) the common mean and variance between every pair of variables, but does not maintain the variance or correlation coefficients. In particular, it increases the variance of the variables, while reducing the correlation, in absolute value, between each pair of variables, due to the added noise element.

- **Adding adjusted noise**—In this method, we randomize adjusted noise regarding a number of variables. It can be shown that in this method, the correlations between each pair of variables are maintained.

A common problem in adding additive noise is that for high and low values of the variable, noise of the same scale is added. This means that a relatively high value changes only slightly, while a low value changes greatly in relative terms. One way of solving this is to add multiplicative noise, which is proportional to the size of the value in the field, instead of additive noise. This method maintains various characteristics of the fields, such as mean and variance.

- **Micro-aggregation**[10]—A method that reduces the level of information in the field, and is mainly intended for continuous fields. This method can be used on one field or on a number fields simultaneously. It takes one field or a number of fields and divides the records into a number of groups with at least k records in each field. In each group, the value of the fields are replaced with the group average. The principle in the division is to create groups with maximum homogeneity within the group. This method ensures that the file that is distributed contains records that fulfill the k-anonymity requirements. To illustrate on one field, the following is a numerical table that divides the records into groups, each of which contains at least two records where the original values are replaced by the group average.

Record	Old value	New value
1	25	21.5
2	12	9
3	18	21.5
4	10	9
5	105	109
6	99	109
7	5	9
8	122	109

[9] See, for instance, item [8] in the Bibliography.
[10] See, for instance, item [3] in the Bibliography.

- **Post Randomization Method (PRAM)**[11]—A method that is appropriate for categorical fields. Categories within a certain set are replaced through a probability transition matrix with the probabilities of replacing values. We label the categorical variable in the original file as X and the new variable that is created as Y. We assume that the two variables have K similar categories 1,…k. The transition between the X variable and the Y variable is done through a transition matrix with a generic element that is defined for each {K…,1} = ji.

$$P_{i,j} = P(Y=j \,/\, X=i)$$

This expression presents the probability that category i will be changed to category j.

The following is an example of such a matrix that is appropriate for a field with three categories.

In this matrix, we can see that the probabilities on the main diagonal, meaning the probabilities that there will be no change in the category, are the highest. If this matrix is known, we can learn about the characteristics of the original variable, such as its mean and variance.

Matrix with the probabilities that the categories will be replaced

	New category		
Original category	**1**	**2**	**3**
1	**0.8**	**0.1**	**0.1**
2	**0.05**	**0.9**	**0.05**
3	**0.1**	**0.3**	**0.6**

- **Creating a file with synthetic data**—A synthetic file is a file that contains data that are different from the original file, but which is built in a way that proximately maintains the statistical characteristics of the file, such as the marginal distribution of the fields and the correlations between the fields. Even though this method is not preferred by researchers, since they generally prefer access to real data, it can serve to calibrate researchers' models, and conduct trial and error in the absence of access to real data, such as for statisticians conducting anonymization who need a file with similar characteristics.

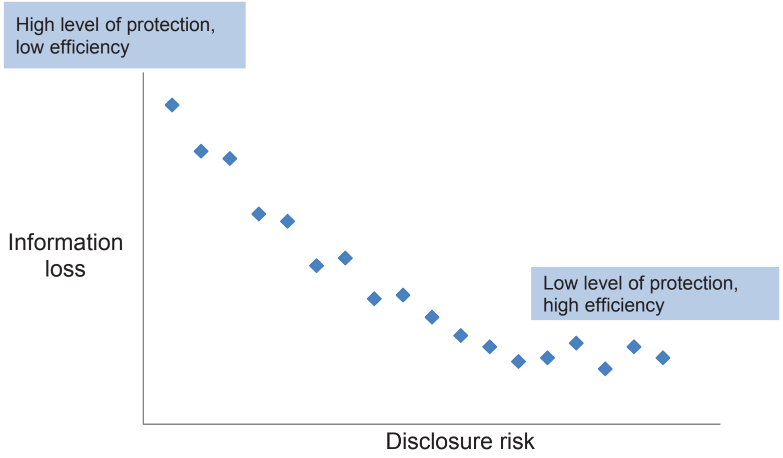
Even though all of the observations are different, this method does not always provide full protection for the file. Similar to a situation in which noise is added to data, it is common to assess the risk in a synthetic file—to what extent can there be record linkage between it and the original file.

[11] See, for instance, items [6] and [2] in the Bibliography.

Stage 5: Assessing disclosure risk in the file and maintaining information efficiency

Maintaining information efficiency and minimizing risk—The objective of the anonymization process is to make a protected file of data accessible so that it embodies a low risk of identification of the individuals, while at the same time, subject to that limitation, maintaining maximum information in the file (information efficiency/usability). There is a tradeoff between the level of information protection and its usability. The higher the level of protection, the greater the information loss (Figure 2).[12] The objective is to find the methods that will lead to the optimum tradeoff given the importance of information use and the damage that may be caused from identification. There are a number of methods for measuring the maintenance of information efficiency in the file, including a direct comparison between the data in the original file and the data after anonymization, and a comparison of calculated statistics (average, standard deviation, and so forth) between them.

Figure 2
Information Efficiency vs. Risk



[12] See item [4] in the Bibliography.

4. CONCLUSION

The Information and Statistics Department uses various complex methods, described above, to anonymize itemized data in a variety of content worlds for users of the information. An effective anonymization process protects the itemized data, while also maintaining the usability of the information even after some of it is lost. The extent of anonymization is determined in accordance with information disclosure scenarios that we want to protect against. Building these scenarios is a complex process that requires expertise in content and also takes into account the existence of complementary databases that are available to users and enable cross-referencing of information and identification of the individuals.

In an era in which information analysis is based more and more on powerful databases of itemized data, the Bank of Israel will have to continue conducting complex anonymization processes in order to allow for freedom of information for policy and economic research needs, while at the same time maintaining the confidentiality of the itemized information as required by law.

BIBLIOGRAPHY

[1] Anco Hundepool, Josep Domingo-Ferrer, Luisa Franconi, Sarah Giessing, Eric Schulte Nordholt, Keith Spicer, and Peter-Paul de Wolf (2012), Statistical Disclosure Control, First Edition.

[2] Dalenius, T. and S.P. Reiss (1978), “Data-Swapping: A Technique for Disclosure Control”, Proceedings of the ASA Section on Survey Research Methods, pp. 191–194. American Statistical Association, Washington, DC.

[3] Defays D. and P. Nanopoulos (1993), “Panels of Enterprises and Confidentiality: The Small Aggregates Method”, Proceedings of the 92nd Symposium on Design and Analysis of Longitudinal Surveys, pp. 195–204. Statistics Canada, Ottawa.

[4] Duncan G., S. Keller-McNulty and S. Stokes (2001), “Disclosure Risk vs. Data Utility: The R-U Confidentiality Map”, Technical Report LA-UR-01-6428, Los Alamos National Laboratory, Statistical Sciences Group, Los Alamos, New Mexico.

[5] Gehrke J., D. Kifer, A. Machanavajjhala, and M. Venkitasubramaniam (2006), “L-diversity: Privacy Beyond K-Anonymity,” 22nd International Conference on Data Engineering (ICDE’06), Atlanta, GA.

[6] Gouweleeuw J.M., P. Kooiman, L.C.R.J. Willenborg, and P. P. de Wolf (1997), “Post Randomization for Statistical Disclosure Control: Theory and Implementation”, Technical Report, Statistics Netherlands. Research paper no. 9731.

[7] Samarati P. (2001), “Protecting Respondents’ Identities in Microdata Release” IEEE Transactions on Knowledge and Data Engineering 13(6), pp. 1010–1027.

[8] Sullivan G.R. (1989), “The Use of Added Error to Avoid Disclosure in Microdata Releases”, Ph.D. Thesis, Iowa State University.