Irving Fisher Committee on
Central Bank Statistics

BANK FOR INTERNATIONAL SETTLEMENTS

Ninth IFC Conference on "Are post-crisis statistical initiatives completed?"

Basel, 30-31 August 2018

The establishment of a central credit register at the Bank of Israel and its statistical disclosure control processes[1]

Ariel Mantzura,
Bank of Israel

[1] This paper was prepared for the meeting. The views expressed are those of the authors and do not necessarily reflect the views of the BIS, the IFC or the central banks and other institutions represented at the meeting.

The Establishment of a Central Credit Register at the Bank of Israel and its Statistical disclosure Control Processes

Ariel Mantzura

Abstract

The Information and Statistics Department at the Bank of Israel collects and produces economic statistics and manages databases that contain granular data in various fields: the equity market, foreign exchange, banking, credit, and more. The Bank of Israel is now intensively engaged in the building of a central credit register containing granular and personal information regarding the credit history of individuals, which will serve credit agencies in the building of models for credit scoring. On the basis of this credit register the Information and Statistics Department at the Bank of Israel will manage an anonymized database where the private information is unidentified. This system is built for in-house use in order to support some of the tasks assigned to the central bank by law. This work will describe the credit register's statistical disclosure control processes.

1. INTRODUCTION

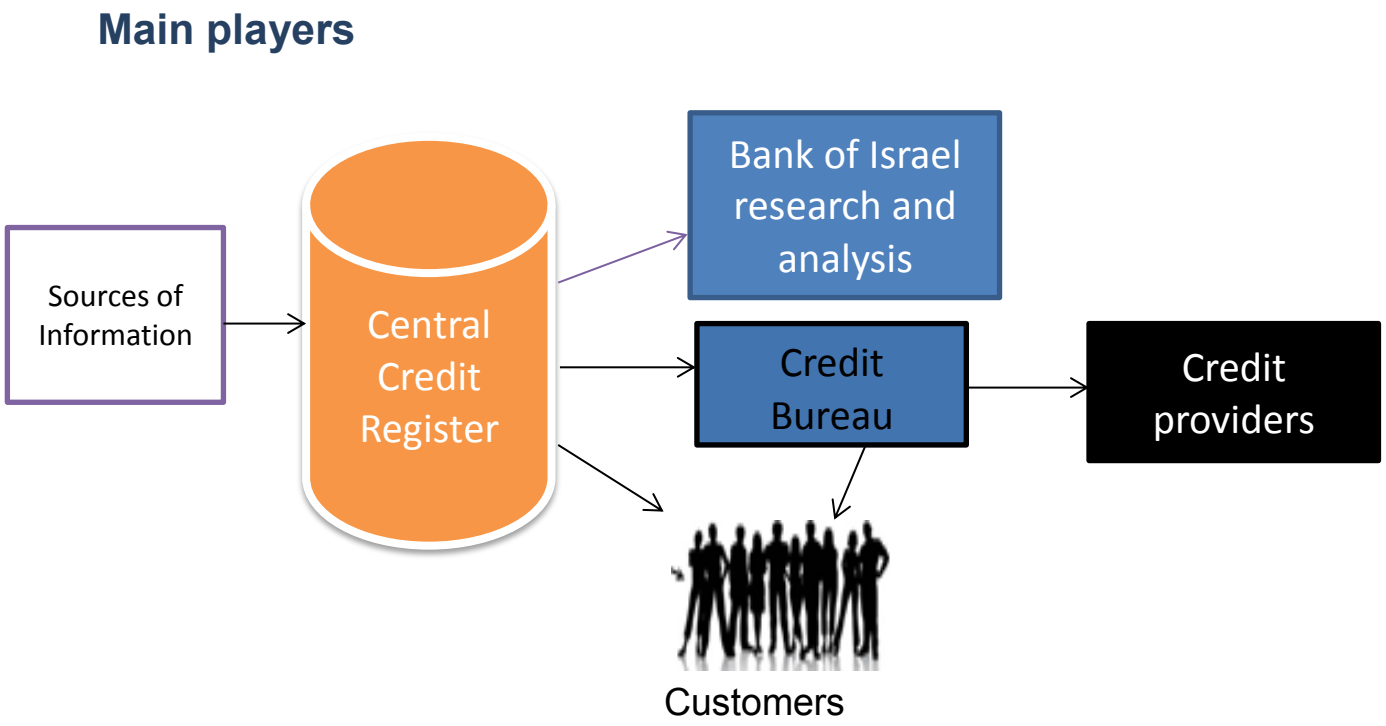
Following the Global Financial Crisis of 2008, central banks, including the Bank of Israel, began managing macroprudential policy, the aim of which is to identify systemic risks at the formative stage and to advance actions that will deal with them and limit their effect on the financial stability of 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

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

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 this database, while also maintaining its confidentiality, the Bank of Israel is designing a process called "anonymization". 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. 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 the credit agencies, subject to the privacy protection restrictions and maintaining commercial confidentiality.

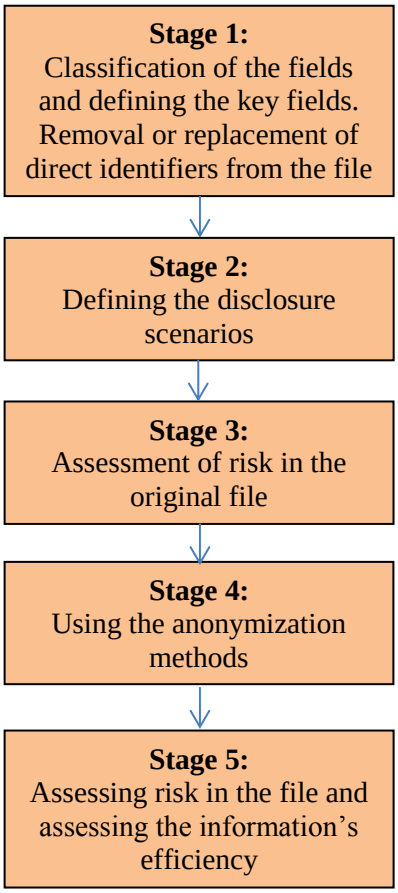
A database containing itemized information naturally includes information that directly identifies the individual—a field that on its own exposes the identity of the individual even without needing additional information located in other fields. Examples of this include the identification number and 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 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. 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.

2. DESCRIPTION OF THE STAGES IN THE ANONYMIZATION PROCESS

Flowchart of the anonymization process



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**[2]—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.

[2] See, for instance, [7].

- **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. An example of such a field is a person's income.

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

- **Categorical fields**—Fields that include a finite number (generally a low number) of categories or values.
- **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[3] 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.

In the context of the credit register, a designated committee composed of domain experts, statisticians, the supervisor of privacy protection and legalists determined together the disclosure scenarios regarding the internal uses of the credit register. These disclosure scenarios sum up to the construction of a three column table with the following structure (example):

[3] See, for instance, [7].

Database that can be cross-referenced	Type of data in credit register that can be cross referenced	Who has access to both databases?
Personal information	• Direct identifiers • Exact numeric values	• Credit register users
Employees file	• Income • Payment to Income • Statistical area	• Research, Data and Statistics, IT
Real-estate transactions file	• Statistical area • Loan to Value • Loan date	• Research, Data and Statistics, IT

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 likelihood that it will be possible to connect a certain record in a file and 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**[4]—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 likelihood of risk 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.
  - **L-diversity requirement**[5]—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 l-diversity requirement is that in all possible combinations there should be at least *l* 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.

- **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

[4] See, for instance, [7].
[5] See [5].

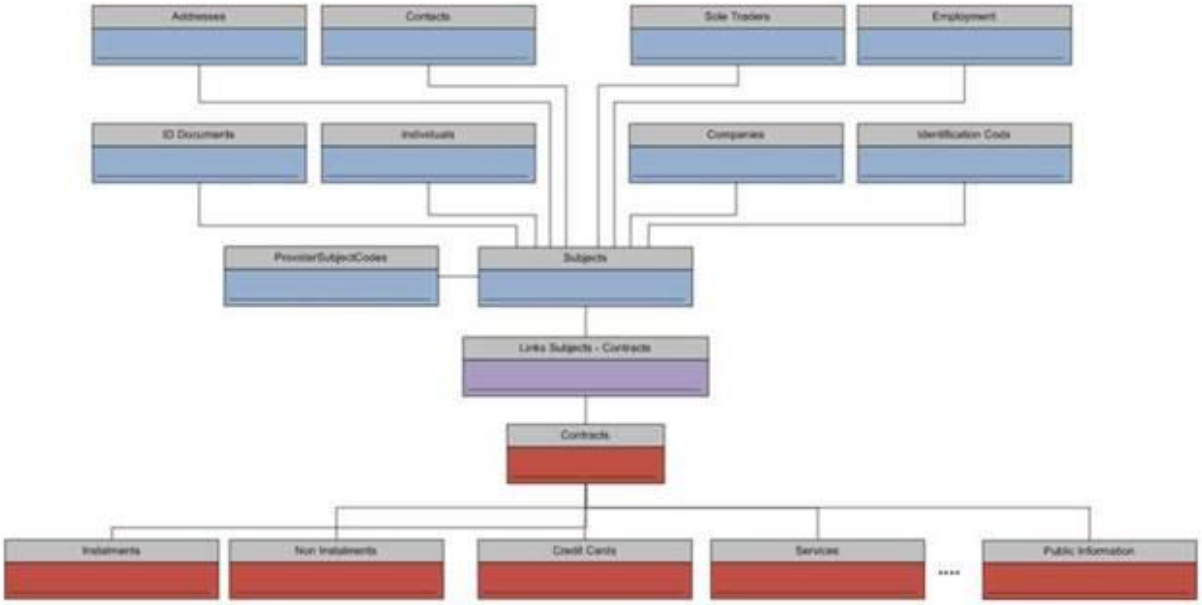
assess the risk in these variables based on the extent to which record linkage is possible between the file where we changed the data on continuous variables, such as by adding noise, and the original file.

- **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 likelihoods of identification of the records in the file. An example of such an index is the total likelihood of identification in the file, which is equal to the incidence of the number of identities in it.

Structure of database

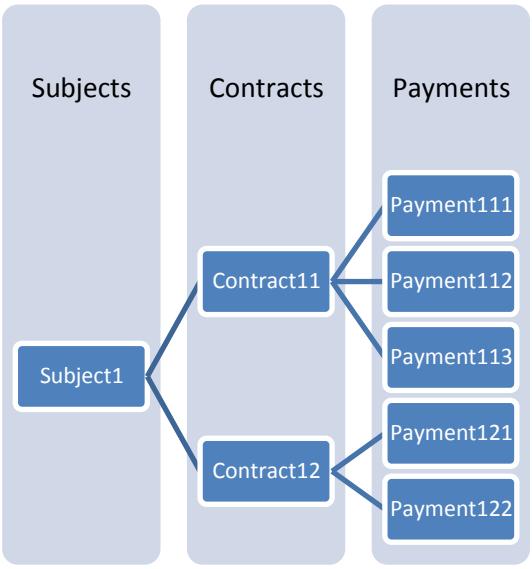
Assessing the risk of the database via the concept of k-anonymity requires a 2- dimension structure of the data, i.e. rows and columns. However, the primary structure of the database consists of linked tables which does not enable assessing risk in the k-anonymity sense.

Structure of linked tables



Evaluating the risk via the concept of k-anonymity requires us to flatten the linked tables structure to a 2-dimensional table with rows and columns. The columns must of course contain key fields. The following figure is an example of a flat structure.

Flat structure



Stage 4: Using the anonymization methods

The following is an outline of a number of common anonymization methods that were considered (among others) in the anonymization of the credit register:

- **Global recoding**—a method that lowers 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. 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 an individual case of global recoding, and deals with the ends 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.
- **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 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.

- **Adding noise (additive)**[6]—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 (unadjusted)**—In this method, unadjusted noise is added to a particular field, which we will label as X, as follows:

    $$Z=X+\varepsilon$$

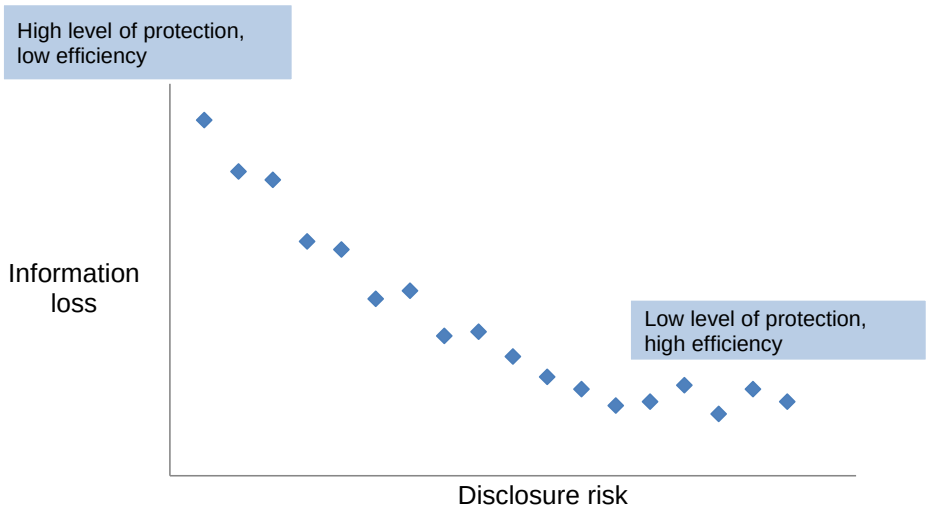
    where ε is a vector of noises broken down numerically and unadjusted (white noise). It can be shown that this method maintains (proximately) the common incidence and variance between every pair of variables, but does not maintain the variance or correlation coefficients. In particular, it increases the 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.

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, it maintains 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).[7] 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.

[6] See, for instance, [8].
[7] See [4]

Figure 2
Information Efficiency vs. Risk



3. CONCLUSION

The Information and Statistics Department uses various complex methods, described above, to anonymize itemized the data of the central credit register. 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.

Irving Fisher Committee on
Central Bank Statistics

BANK FOR INTERNATIONAL SETTLEMENTS

Ninth IFC Conference on "Are post-crisis statistical initiatives completed?"

Basel, 30-31 August 2018

The establishment of a central credit register at the Bank of Israel and its statistical disclosure control processes[1]

Ariel Mantzura,
Bank of Israel

[1] This presentation was prepared for the meeting. The views expressed are those of the authors and do not necessarily reflect the views of the BIS, the IFC or the central banks and other institutions represented at the meeting.

The establishment of a Central Credit Register at the Bank of Israel and its Statistical Disclosure Control processes

Ariel Mantzura
Bank of Israel

August 31, 2018

Credit Data Law - Objectives

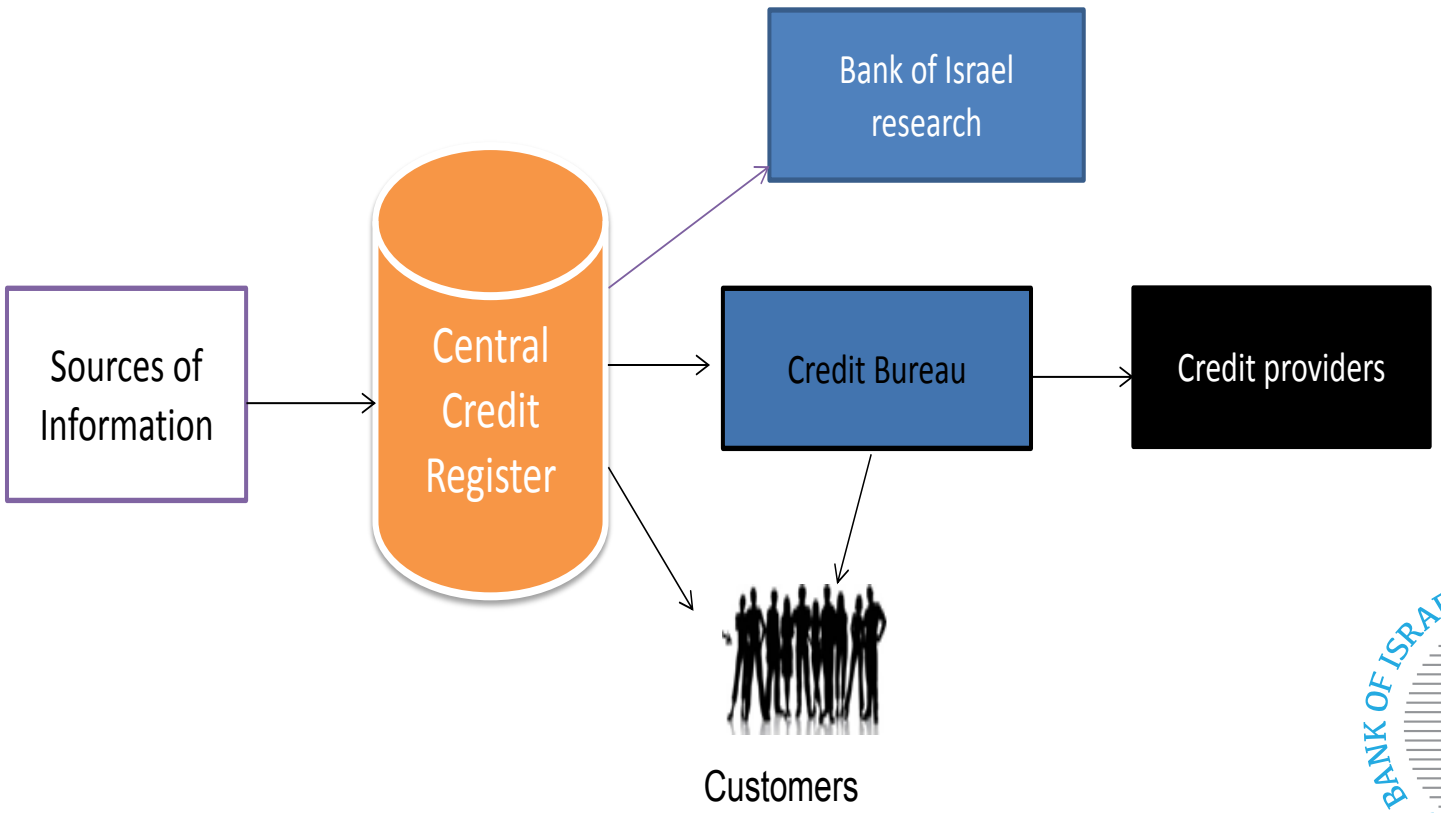
The objective of this law is to establish an overall arrangement for sharing credit data...for the following purposes:

- Enhancing competition in the retail credit market.
- Expanding access to credit.
- Reducing of discrimination in the granting of credit and of economic gaps.
- Creating an anonymous database for use by the Bank of Israel in carrying out its functions.

Backround - Protection of Privacy Law

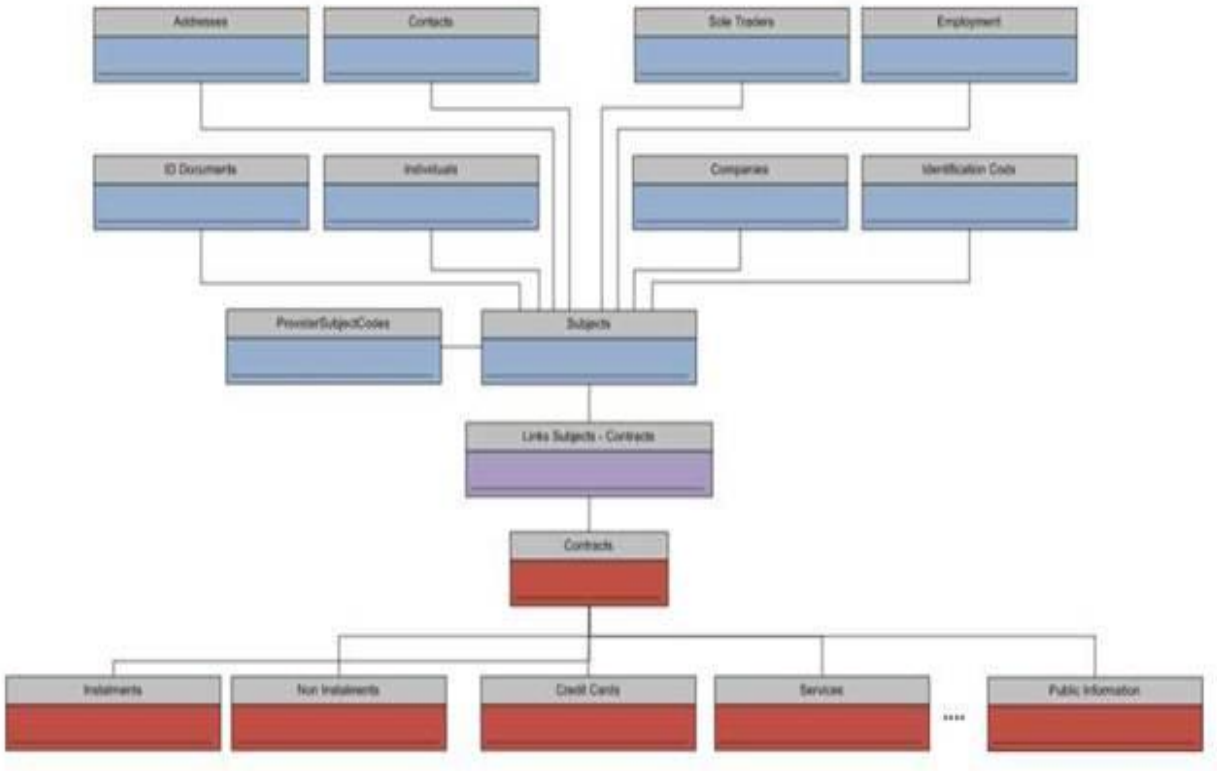
- In order to allow access to such information within the organization or outside it, the Protection of Privacy Law requires that the confidentiality of the information be maintained, as the information relates to individual persons.
- In addition, the law requires that the commercial confidentiality of business entities be maintained, a complex task, particularly when dealing with financial information that is sometimes characterized by high concentration.

Backround - Central Credit Register



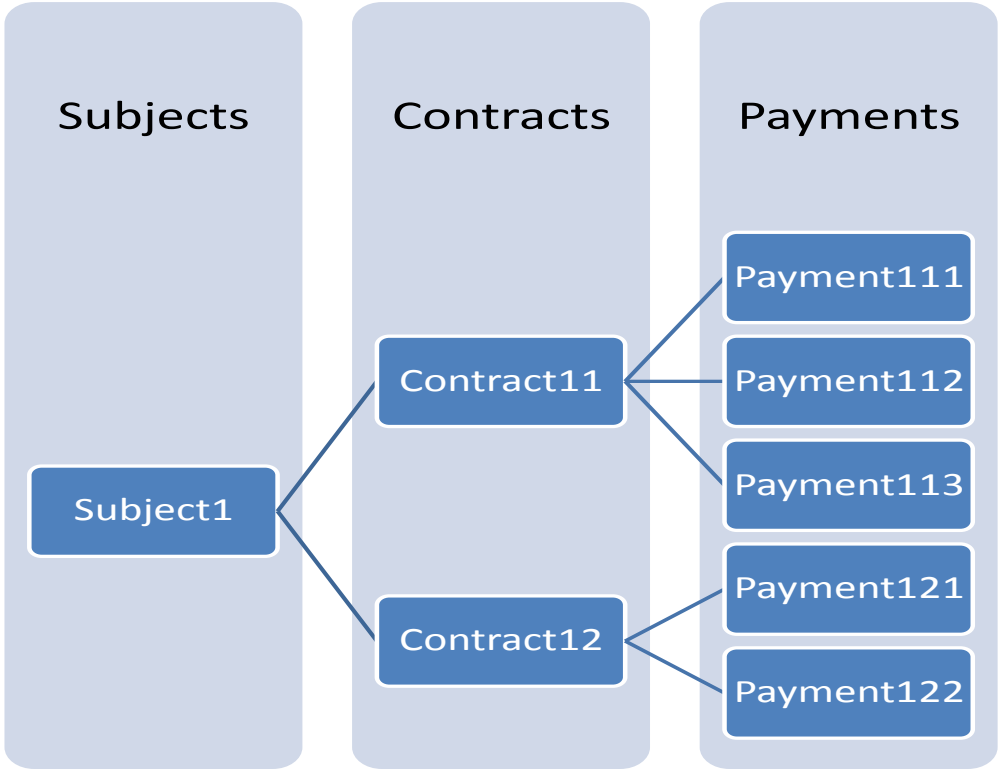
Backround - Credit register structure

Linked tables structure

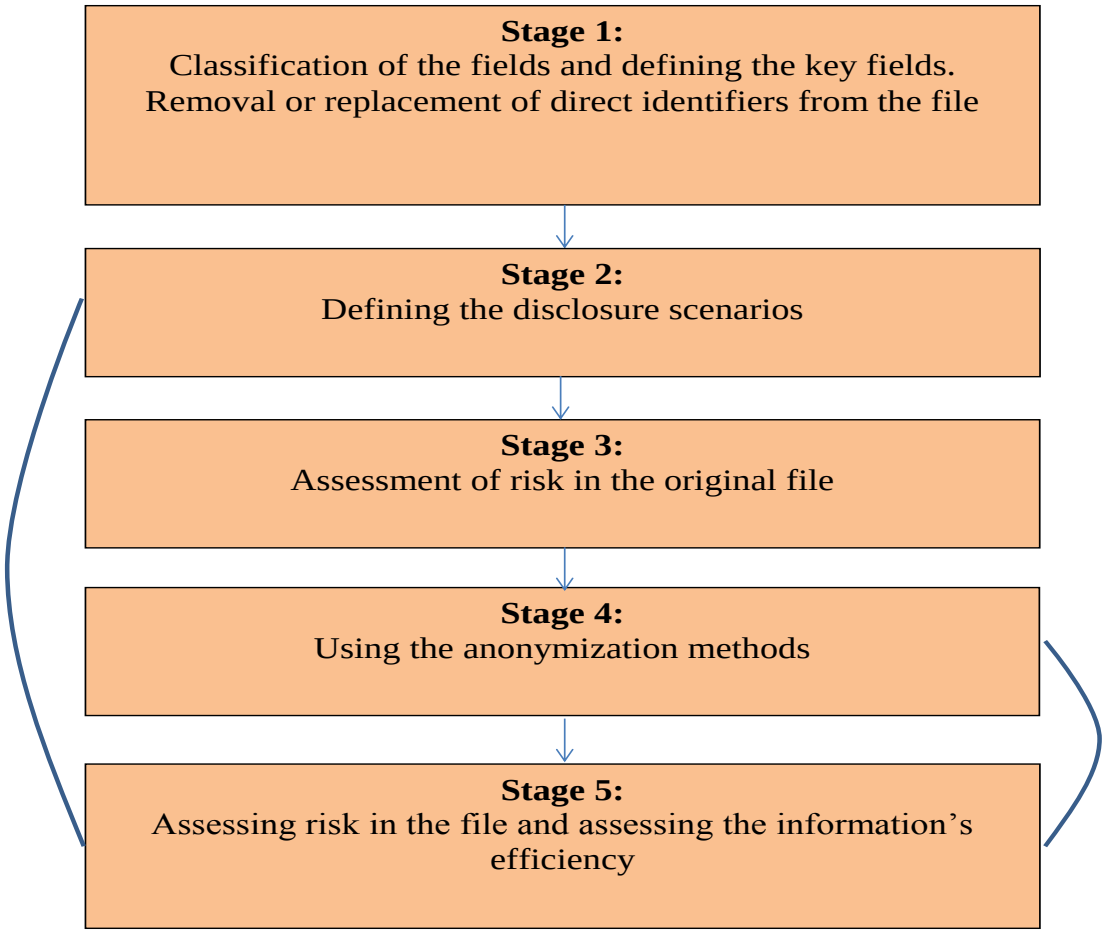


Backround - Credit register structure

Flat structure



Flow chart of anonymization process



Defining disclosure scenerios

- Disclosure scenarios 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.
- 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.

Defining disclosure scenerios

- 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.
- The disclosure scenarios are determined with the help of experts in the relevant content worlds.

Defining disclosure scenerios

Database that can be cross-referenced	**Type of data in credit register that can be cross referenced**	**Who has access to both databases?**
Personal information	• Direct identifiers • Exact numeric values	Credit register users
Mortgage file data	•	•
Employees file	•	•
Real-estate transactions file	•	•

Defining disclosure scenerios

- The disclosure scenarios can be less or more severe than the objective information disclosure possibilities.
- The disclosure policy 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.

Defining disclosure scenerios

- It is common to distinguish between **scientific use files** and **public use files**.
- **scientific use files (SUF)** 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.
- **public use files (PUF)** have no restriction or control. The policy regarding the information files issued to the public is generally very strict.

Assessing the risk of disclosure in the file

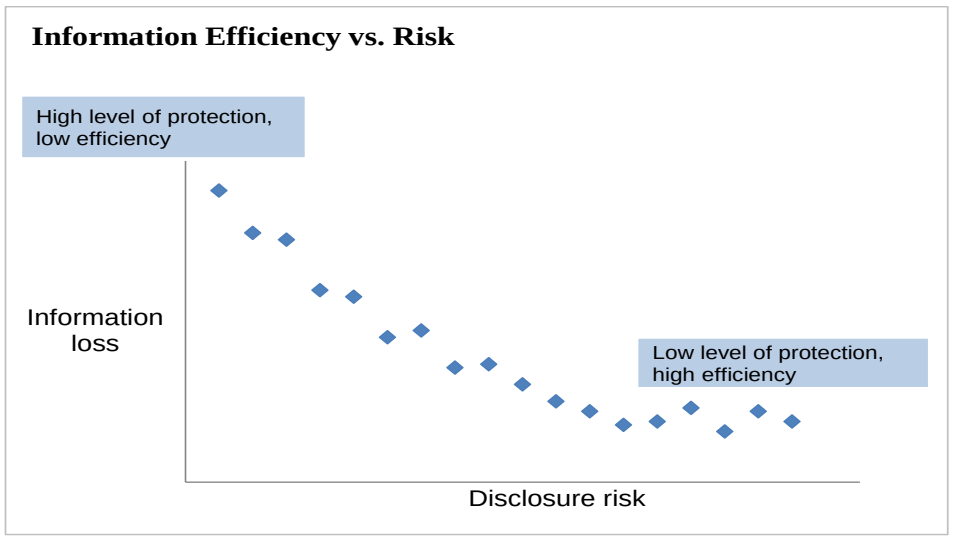
- There are two common requirements:
- **K-anonymity** - a requirement that in each combination of categorical key fields in groups that are defined in the disclosure scenario, there shall be at least k records with the same combination.
- **l - diversity** - The l-diversity requirement is that in all possible combinations there should be at least l different values.

Information efficiency vs 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.
- At the same time, subject to that limitation, it maintains maximum information in the file.
- There is a tradeoff between the level of information protection and its usability.
- The higher the level of protection, the greater the information loss.

Information efficiency vs minimizing risk

Figure: Risk Utility Map



Thank you!