A graphical framework to evaluate risk assessment and information loss at individual level

Giovanni Seri*
* Istat – Italian National Statistical Institute, 00184 Rome, Italy, seri@istat.it

Abstract: When dealing with statistical disclosure control (SDC) problems two aspects have to be considered. Firstly, a rule based on a measure of the risk of disclosure has to be adopted in order to decide if a certain release of data is safe or unsafe. Secondly, protection methods have to be applied to reduce the risk of identification when the release of data is classified as unsafe. The performance of a protection method is usually measured in terms of 'risk of disclosure' and 'information loss'. In this work we present a graphical framework named 'confidentiality plot' for the evaluation of risk of disclosure at individual level. For some extent the tool can be jointly used to evaluate risk of disclosure and information loss.

1. Introduction

National Statistical Institutes (NSIs) facing the problem to release micro data file for research (MFR) usually adopt some statistical methods to preserve confidentiality of statistical units. We consider that breach of confidentiality is produced if a statistical unit is re-identified and the value of some sensitive variables is disclosed. A measure of the re-identification risk is then needed in order to classify data as 'at risk' or 'safe'. When data are at risk some disclosure limitation methods have to be applied in order to reduce the level of re-identification risk under a pre-defined threshold assumed as acceptable.

The definition of a disclosure scenario is a first step towards the development of a strategy for producing a "safe" MFR. Indeed, a scenario synthetically describes (i) which is the information potentially available to the intruder, and (ii) how the intruder would use such information to identify an individual: i.e. the intruder's attack means and strategy. We refer to the information available to the intruder as an *External Archive* containing directly identifying variables (name, identification number, etc.), and some other variables about individual or enterprises. Some of these further variables are assumed to be available also in the data set we want to protect. We assume that the intruder tries to match the information in the individual archive with that in the micro data file (for instance through record-linkage). We refer to these matching variables as *key or identifying variables*. Therefore, the selection of key variables is crucial.

Statistical limitation methods suggested in literature can be classified on the base of their impact on the data in two categories (Domingo-Ferrer and Torra, 2001a):

- methods based on data reduction;
- methods based on data perturbation.

Methods based on data reduction aim at increasing the number of individual in the population sharing the same or similar identifying characteristics presented by the investigated statistical unit in order to avoid presence of unique or rare recognizable individuals. Perturbation methods, on the contrary, achieve data protection from a twofold perspective. On one hand, if the data are modified, re-identification by means of record linkage or matching algorithms is made harder and uncertain; on the other hand, even when an intruder is able to re-identify a unit, he/she cannot be sure that the data disclosed are consistent with the original data. In this work we consider methods based on data reduction for MFR from social survey and methods based on data perturbation for MFR from business survey. A different approach based on synthetic micro data is out of the scope of this work (see for example, Abowd and Lane, 2004; Polettini, 2003).

Summarising, safety of a record characterised by certain values of the identifying variables is generally evaluated by two aspects: (i) the number of individual sharing similar identifying characteristics in the population and (ii) the difference between the data released and the original data. In this paper we present a graphical framework where statistical units are plotted with coordinates representing these two aspects. We call this framework as 'confidentiality plot'. Rules to classify units as 'safe' or 'at risk' can be represented in the graph making identification of units 'at risk' easier. Confidentiality plot can be used to assess the performance of a disclosure limitation method on the base of the individual re-identification risk and, for some extent, also on the base of the level of information loss.

The connection between the "confidentiality plot" and the R-U confidentiality map of Duncan et al. (2001a, 2001b) is clear. R-U confidentiality map were introduced to compare the performances of different disclosure limitation methods. The main difference is that a point in the R-U confidentiality map represents a disclosure limitation techniques (e.g. a given data release), whereas a point in the confidentiality plot represents individual data. See also Sebé *et al*. 2002.

Details on available re-identification criteria based on the rareness (uniqueness) of a unit in the population used in the case of social micro data can be found in Skinner and Holmes (1998), Fienberg and Makov (1998), Benedetti and Franconi (1998), Franconi and Polettini (2004). As regards enterprises micro data, the approach suggested in the literature to assess the risk of disclosure refers to record linkage procedures. Domingo-Ferrer and Torra (2001b) report on "distance based record linkage" (Pagliuca and Seri, 1999), "probabilistic record linkage" (Winkler, 1998, Yancey et al., 2002) and "interval disclosure" keeping into account different sets of key variables or parameters.

In the next Section we outline the framework used for the risk assessment based on confidentiality plot. Some empirical results on business perturbed micro data are presented in Section 3. Some conclusions and future perspective are given in Section 4.

2. Confidentiality plot

We consider the micro data set to be protected as a matrix A with n rows representing units and m+s columns representing the m key variables (x_j, j=1,…,m) and the s confidential variables (c_r, r=1,…,s) respectively:

A=(X,C), where X={$x_{i,j}$, i=1,…,n; j=1,…,m} and C={$c_{i,r}$, i=1,…,n; r=1,…,s}. (1)

The matrix C usually is not involved in the risk assessment and will be ignored in the following. We can assume that the application of protection methods consists in replacing X with a different matrix Y={$y_{i,j}$, i=1,…,n; j=1,…,m}.

We firstly consider the case of social data in which identifying variables are mainly categorical or can be treated as categorical (the variable "Age" for example). Let y be the combination of identifying variables presented by a given record 'i' in the micro data file with f_y and F_y respectively the frequency of the same combination in the micro data file and in the population. Suppose the re-identification risk is $r_y=1/F_y$, that can be interpreted: a statistical unit represented as a combination of value of some identifying variables is "at risk" if the same combination is "rare" in the population or, equivalently, the higher is F_y the lower is the risk associated to the combination y. As a disclosure scenario we consider the external archive being a complete and reliable register of the population containing: place of residence (at municipality level), date of birth, sex and marital status. The intruder strategy is to link records presenting the same combination of key variables in both the external archive and in the MFR.

As an instance, we might consider a woman, born in 1973, October the 23rd, resident in Montemignaio (a small Italian municipality with less then 1000 inhabitants) and divorced. There is high probability that it is a unique case in the population so that the intruder can easily recognize her, and disclose other information provided by survey micro data. In such a case, protection methods such as global recoding (Willenborg and de Waal, 2001) produce reduced information providing, as an example: (i) age in class instead of date of birth; (ii) region instead of municipality of residence. Of course, F_y relative to the combination (woman, 30-34 years old, residents in Tuscany and divorced) increase as the risk decrease. Moreover, consider a woman, 15-19 years old and widow classified according to the previous recoding. Nevertheless, there is high probability that F_y is very low and the unit recognizable in the population. In this case a method as 'local suppression' (Willenborg and de Waal, 2001) reduce the information replacing a variable (for example marital status) with missing value. The combination (woman, 15-19 years old) replace the original one and the corresponding F_y is higher.

Under such a scenario the risk is measured on the base of the number of individual in the population sharing the same combination of identifying variable of each record in the file. Similar reasoning can be made considering f_y instead of F_y or replacing F_y with a proper estimate if the true value is unknown (Franconi and Polettini, 2004).

As regards business micro data most of the information collected takes the form of quantitative variables with skew distributions. Such variables are often representative of enterprise size and are extremely identifying. For example, information about turnover can lead to the identification of a very large and well-known enterprise in a particular class of the NACE classification of economic activity. This means that, even though they are not always publicly available, quantitative variables have to be considered as key variables. The practical consequence of this is that all units are unique (rare) with respect to a small set of quantitative variables. Moreover, in many cases, populations of enterprises are sparse and firms are easily identifiable simply by their economic activity and geographical position. Finally, other a priori information such as knowledge about the survey design can be used to identify an enterprise, see Cox (1995).

As a consequence, many of the protection techniques specifically proposed for business micro data aim at perturbing the original data in such a way that enterprises are not recognisable. As previously stated, perturbation methods achieve data protection from a twofold perspective. On one hand, if the data are modified, re-identification by means of record linkage or matching algorithms is made harder and uncertain; on the other hand, even when an intruder is able to re-identify a firm, he/she cannot be sure that the data disclosed are consistent with the original data. Of course, this latter aspect has to be balanced with the need to make the information content of perturbed data as similar as possible to that of the original data in order to preserve the quality of statistical results.

We define the disclosure scenario for a NSI assuming that the external archive available to the intruder coincides with the original file, X. We consider this disclosure scenario as the "worst" for a NSI because it is implicitly assumed that the intruder knows that: (i) the target enterprise is included in the released file and (ii) there are no differences between the original data and the external archive due, for example, to classification errors. In this paper we assume that the set of identifying variables consists in Turnover and Number of employees.

As data are perturbed, a strategy of attack based on exact record linkage will probably results in failure of matching. Therefore we assume that the intruder will mark as possible links those records of the external archive which are similar to the target with respect to the set of key variables. This leads us to consider the concept of neighbourhood of a released record. For each record y in the released file, we denote as neighbour of y any unit x in the original sample (the external archive) that is "similar" to y. The level of protection ensured by each perturbed unit will depend on the number of neighbours we can attach to it.

In Figure 1 (a) two triangles - with coordinates (7.3,15.5) and (3.4,9.6) - representing two units treated with the same amount of perturbation are plotted against the original values (the external archive), logarithmic scale is used. In plot (b) and (c) the

positions of those two perturbed units is zoomed in. Figure 1 (b) represents the situation where an outlier in the original data is weakly perturbed by the protection method. As the protected record (the triangle) is very close (similar) just to a single isolated point, an intruder trying to compare the released record with the data in his/her archive will have great confidence that the link between two such points is a true link. Figure 1 (c) shows the protected record confused in a crowded cloud of points, and of course this makes it harder to identify the correct link because a high number of enterprises share similar values of turnover and number of employees.

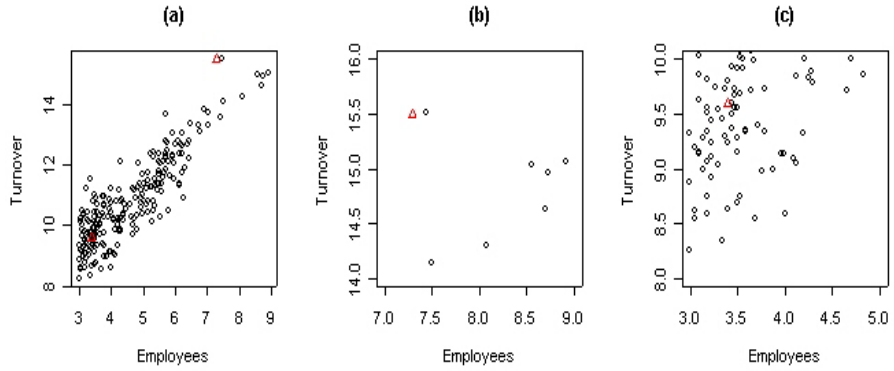

Figure 1 Quantifying the extent of protection by the number of neighbours

We then argue that the amount of perturbation induced in the data and the number of neighbours, can be jointly exploited to assess the protection of a record. In particular, a graphical tool connecting the above mentioned aspects can be introduced. We denote by "confidentiality plot" a graph in which protected data can be represented with coordinates the "number of neighbours" (horizontal axis) and the "amount of perturbation" (vertical axis). A general scheme for the confidentiality plot is presented in Figure 2.

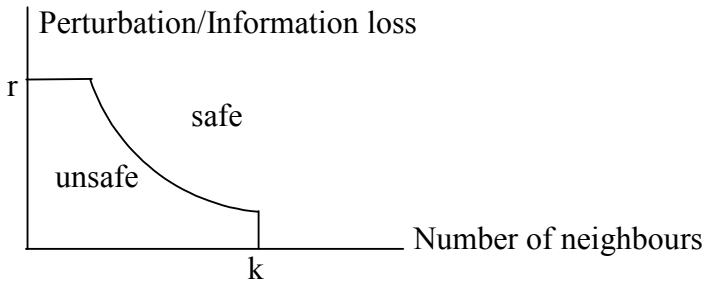

Figure 2 A general scheme for confidentiality plot

The threshold "r" means that the released value is safe if it is distorted over the r% of the original value, whatever the risk of re-identification. On the other hand, the threshold "k" means that if the perturbed value is close to more than k units in the

population, then no perturbation is required to protect this value. The curve represents the trade-off between these two aspects: the more the released value is confused in the population, the less the perturbation that is required, and vice versa. The area under the curve is defined "unsafe" zone, because points in this area represent records that are not protected enough.

The position of each point in the confidentiality plot with respect to the vertical axis can also be interpreted as an index of the quality of representation. In other words "information loss" and "perturbation" induced in a single record are equivalent labels for the vertical axis (see Section 2.1).

In the case of social data described above the confidentiality plot can result as a vertical line in k being the threshold of safety fixed for a given frequency F_y=k. If no perturbation methods are applied the vertical axis is not more representative of safety of records. Nevertheless, a proper measure of information loss at individual level can be defined and represented on the vertical axis of the confidentiality plot.

2.1 A way to measure information loss/perturbation and count neighbours

In this section we outline a proposal to measure the amount of perturbation (information loss) and a way to count the numbers of neighbour for each record.

Firstly we assume that 'perturbation' can be represented individually for each record by the difference between the original data and the corresponding perturbed data independently from the SDC method used. The more this difference is small the better the record is represented in the MFR and the lower is the information loss. In other words, both information loss and perturbation can be suitably represented by a measure of the distance, e.g. the Euclidean distance, between the perturbed and the original record.

The aim is to measure the error that is to be accepted by a user accessing the released data in place of the original data. Denoting by y the vector of key variables for a generic record in the released file, and by x the corresponding vector of true values, we compute as the relative error:

$$\text{Information loss} = ||y-x||/||x||. \quad (2)$$

Clearly, the measure of information loss in (2) is also a measure of the perturbation induced in the data, as it represents the distance of the released value from the truth.

Secondly, in order to define a neighbourhood to compute the number of neighbours of each record, a measure of similarity between units is needed. Moreover, a linkage procedure is based on a comparison between a released record y and a record x, say d(x,y). As the data are perturbed and variables are numerical, d(x,y) can be defined as a distance. The distance induces a variable Z defined as: z=d(x,y). Clearly, the definition of "similarity" and hence of neighbourhood is arbitrary as it depends on the intruder's strategy of attack. All the possible pairs (x,y) are in the product space

XxY. Let M be the set of pairs corresponding to correct links and U the set of nonlinks:

$XxY=M \cup U$.

The two distributions:

$$m(z) = P(Z = z \mid (x,y) \in M) \quad \text{and} \quad u(z) = P(Z = z \mid (x,y) \in U) \tag{3}$$

are the basic ingredients of probabilistic record linkage and their estimation is the main issue in the record linkage literature. As in our framework the two archives involved in the procedure are completely known, we can assign without uncertainty each pair to link/nonlink. Therefore the disclosure scenario assumed in this work allows us to compute the two distributions in (3).

In order to compute the number of neighbours of each record, according to the record linkage approach described above, we define the comparison variable Z as the relative euclidean distance between a released record y and each record x in the external archive:

$Z=||y-x||/||y|| \quad \forall\, y \in Y; \forall\, x \in X.$

For each y to be released, the number of neighbours is computed as the number of original records $x \in X$ such that z is lower than a threshold δ (we remind that X is also the external archive available to the intruder):

$$\text{Number of neighbours of } y = \#\{ x \in X : z < \delta\}. \tag{4}$$

We then denote as "neighbourhood" of y:

$N(y)=\{ x \in X : z < \delta\}$.

As in the probabilistic record linkage we consider the probability of the type 1 error, i.e. the probability of designating a pair as a link when it is not. The two densities of the distributions m(z) and u(z) in (3) can be estimated over the set of pairs $(x,y) \in M$ and $(x,y) \in U$ respectively. We remind that the two sets M and U are completely known for the given disclosure scenario. Assuming that lower distances are likely to be measured in the occurrence of a true link, we have:

$Pr(z < \delta | (x,y) \in U)=\alpha$

where α is the acceptable level for the type 1 error and δ, the critical distance, is fixed accordingly. In practice, the neighbourhood of y consists of all the units $x \in X$ that, for given α, are not rejected as nonlinks according to the probabilistic record linkage procedure. Choosing a smaller α turns into reducing the number of neighbours of the released record. As an alternative, it is possible to use the Fellegi-Sunter (1969) approach to record linkage.

3. Experimental results

Data used in this work come from the Community Innovation Survey (CIS) and are treated with the statistical disclosure control techniques proposed in Polettini *et al.* (2002). We explore the assessment of the level of protection guaranteed by the method using the above defined confidentiality plot on 157 enterprises from the CIS sample classified in the division 18 of the NACE nomenclature. We assume the disclosure scenario defined by an external archive equivalent to the original data file (see Section 2) containing turnover and number of employees as key variables.

Figure 2 shows the confidentiality plot when α=0.05 (see Section 2.1). For each record y in the micro data file under investigation, a point is plotted on the graph with coordinates: the number of neighbours (horizontal axis) and the information loss/perturbation (vertical axis). Squares identify 5 cases for which the nearest-neighbour is the correct link, that is when for a given record y the pair $(x,y)\in M$ and x is the nearest neighbour of y. Crosses identify 33 cases for which the correct link is in the neighbourhood, that is when for a given record y the pair $(x,y)\in M$ and $x\in N(y)$. A filled square highlights the unit presenting the highest Turnover in the original data, which in most cases is the most easily re-identifiable.

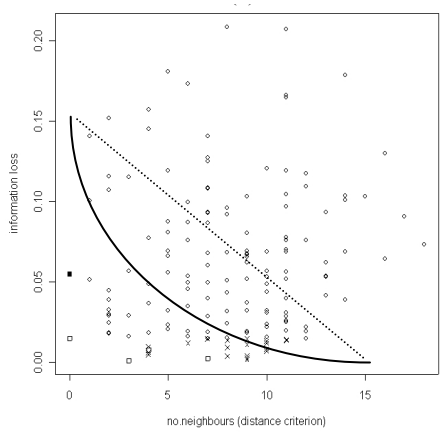

Figure 3 Confidentiality plot for perturbed micro data: α=0.05

The two lines in the plot represent two different hypothesis of confidentiality policy (no real situation is taken into account). The curves consist of combinations of values of 'perturbation' and 'number of neighbours' joining the two points representing approximately the following rules for safety of a record: (i) a 15% of difference (on the logarithmic scale) between the original data and the corresponding perturbed record; (ii) a number of neighbours of 15 and null level of perturbation. The straight line defines stronger rules to preserve confidentiality. Anyway, the plot highlights the presence of outliers (enterprises relatively too large and easily identifiable) and the need for higher protection of the records in the unsafe zone of the plot. Nevertheless, results should be interpreted in the light of the severe disclosure scenario we assumed, particularly: the hypothesis that the external archive coincides with the

original data. It worth noting that most (or all) of the crosses fall down in the unsafe area. For a few units the neighbourhood is empty. These are represented in the confidentiality plot with coordinate Number of neighbours=0. In this case the perturbation applied to records y is higher than the critical distance defining the neighbourhood (δ=0.040838); i.e., any record x is "close enough" to y.

7. Conclusions

Assessing the performance of a disclosure limitation method is a difficult task particularly for business micro data. We have outlined a way to assess graphically by the so called 'confidentiality plot' the level of protection guaranteed by SDC methods at record level. We also introduce the framework for joint evaluation of the disclosure risk and information loss at record level. We argue that in order to assess the disclosure risk of an individual record in a MFR two aspects have to be taken into account: (i) the number of records in the external archive that share the same or similar identifying characteristics of the investigated record; (ii) the difference between the original data and the protected data. These two aspects are assumed as coordinates of each records represented on the confidentiality plot.

Some empirical results have been presented relative to a set of enterprises from the Italian sample of the CIS survey protected by the method presented in Polettini *et al*. (2002). Anyway, the purpose of this work is mainly to present the confidentiality plot as a mean to assess re-identification risk and information loss at record level.

We think that the framework can be adopted in many situation even for tabular data. At this purpose further studies are needed considering as an example: (i) protection methods that results in predictive intervals for numerical variables or (ii) in suppression of cell values in a table (in this case it is often possible to evaluate a feasibility interval for the true value). In both this cases y can be assumed as the interval midpoint. That is, from the intruder's point of view, y as the estimate of x that minimizes the error in (2). Moreover, different intruder's strategies of attack can (o have to) be taken into account. For example: different sets of key variables; different measures of the similarity between units; linkage based on the comparison of ranks for the biggest and therefore more easily identifiable enterprises instead of the nearest neighbour. We will develop these aspects elsewhere.

Acknowledgements

The author would like to thank Luisa Franconi and Silvia Polettini.

The views expressed in this paper are those of the authors and do not necessarily reflect the policies of Istat.

References

Abowd, J. M. & Lane, J. (2004). *New Approaches to Confidentiality Protection: Synthetic Data, Remote Access and Research Data Center.* In J. Domingo-Ferrer and V. Torra (Eds), Privacy in Statistical Database. *Springer-Verlag*, Berlin Heidelberg 2004.

Benedetti, R. and Franconi, L. (1998). *Statistical and technological solutions for controlled data dissemination*. Pre-proceedings NTTS '98, New Techniques and Technologies for Statistics, Sorrento, 1, 225-232.

Cox, L.H. (1995). *Protecting confidentiality in business surveys*. In Business Survey Methods, Cox, B.G., Binder, D.A., Chinnappa, B.N., Christianson, A., Colledge, M.J. and Kott, P.S. (Eds.), New-York: Wiley, 443-476.

Domingo-Ferrer, J., & Torra, V., (2001a). *Disclosure Control Methods and Information Loss for Microdata*. In Confidentiality, Disclosure and Data Access, eds. P. Doyle, J. Lane, J. Theeuwes and L. Zayatz. Amsterdam: North-Holland, 91-110.

Domingo-Ferrer, J., & Torra, V., (2001b). *A Quantitative Comparison of Disclosure Control Methods for Microdata*. In Confidentiality, Disclosure and Data Access, eds. P. Doyle, J. Lane, J. Theeuwes and L. Zayatz. Amsterdam: North-Holland, 111-133.

Franconi, L. & Polettini, S. (2004). *Individual Risk Estimation in μ-Argus: A Review*. In J. Domingo-Ferrer and V. Torra (Eds), Privacy in Statistical Database. *Springer-Verlag*, Berlin Heidelberg 2004.

Duncan, G.T., Keller-McNulty, S.A. & Stokes, S.L. (2001a). *Disclosure risk vs. data utility: the R-U confidentiality map*. Technical Report LA-UR-01-6428, Los Alamos National Laboratory.

Duncan, G.T., Fienberg, S.E., Krishnan, R., Padman, R. & Roherig, S.F. (2001b). *Disclosure risk vs. data utility: the R-U confidentiality map*. In Confidentiality, Disclosure and Data Access, eds. P. Doyle, J. Lane, J. Theeuwes and L. Zayatz. Amsterdam: North-Holland, 135-166.

Fellegi, I.P. and Sunter, A.B., (1969). *A theory for record linkage*. Journal of the American Statistical Association, 64, (1969), 1183-1210.

Fienberg, S.E. and Makov, U.E. (1998). *Confidentiality, uniqueness and disclosure limitation for categorical data*. Journal of Official Statistics, 14, 385-397.

Pagliuca, D. and Seri, G. (1999). *Some results of individual ranking method on the System of Enterprise Accounts Annual Survey*. Esprit SDC Project, Deliverable MI-3/D2.

Polettini, S. Franconi, L. and Stander, J. (2002). *Model Based Disclosure Protection*. In Domingo-Ferrer, J. (Ed.), Inference Control in Statistical Databases: From

Theory to Practice. Lecture Notes in Computer Science, 2316, Berlin: Springer-Verlag, 83-96.

Polettini, S. (2003). *Maximum Entropy Simulation for Microdata Protection*. Statist. Comput., **13**, 307-320.

Sebé, F., Domingo-Ferrer, J., Mateo-Sanz, J.M. and Torra, V. (2002). *Post-Masking Optimization of the Tradeoff between Information Loss and Disclosure Risk in Masked Microdata Sets*. In Domingo-Ferrer, J. (Ed.), Inference Control in Statistical Databases: From Theory to Practice. Lecture Notes in Computer Science, 2316, Berlin: Springer-Verlag, 163-171.

Skinner, C.J. and Holmes, D.J. (1998). Estimating the re-identification risk per record in microdata. Journal of Official Statistics, 14, 361-372.

Willenborg, L. and de Waal, T. (2001). Elements of statistical disclosure control. Lecture Notes in Statistics, 115, New York: Springer-Verlag.

Winkler, W.E. (1998). Re-identification methods for evaluating the confidentiality of analytically valid microdata. Research in Official Statistics, 1, 50-69.

Yancey, W.E., Winkler, W.E. and Creecy, R.H. (2002). Disclosure risk assessment in perturbative microdata protection via record linkage. In Domingo-Ferrer, J. (Ed.), Inference Control in Statistical Databases: From Theory to Practice. Lecture Notes in Computer Science, 2316, Berlin: Springer-Verlag, 135-152.