Book of Short Papers SIS 2018

Editors: Antonino Abbruzzo - Eugenio Brentari

Marcello Chiodi - Davide Piacentino

PUBLISHED BY PEARSON

WWW.PEARSON.COM

First printing, November 2018, ISBN-9788891910233

SIS 2018
49 TH SCIENTIFIC MEETING OF THE ITALIAN STATISTICAL SOCIETY
PALERMO 20-22 JUNE

Contents

SIS 2018
49 TH SCIENTIFIC MEETING OF THE ITALIAN STATISTICAL SOCIETY
PALERMO 20-22 JUNE

1. Preface

This book includes the papers presented at the "49th Meeting of the Italian Statistic Society". The conference has registered 445 participants, 350 reports divided into 4 plenary sessions, 20 specialised sessions, 25 sessions solicited, 27 sessions spontaneous, 2 poster sessions. The high number of participants, the high quality of the interventions, the productive spirit of the conference, the ability to respect the time table, are the main indices of the full success of this conference. The meeting hosted also, as plenary sessions, the ISTAT annual report 2018, and a round table on statistics and job markets. Methodological plenary sessions concerned with ordinal data, the dynamics of climate change and models in biomedicine.
Moreover, two related events were held: Start-up Research (SUR) and Stats Under the Stars (SUS4). The SUS4 event attracted many sponsors of statistical, financial, editorial fields as well as numerous students, not only from Italy but also from abroad (Groningen, Tyumen, Barcelona, and Valencia): 98 students for a total of 25 teams. The SUR was a 2-day meeting where small research groups of young scholars, advised by senior researchers with a well-established experience in different areas of Statistics, was asked to develop innovative methods and models to analyse a common dataset from the Neurosciences.

An innovative approach for the GDPR compliance in Big Data era

Un approccio innovativo per la conformità al regolamento GDPR nell'era dei Big Data

M. Giacalone, C. Cusatelli, F. Fanari, V. Santarcangelo, D.C. Sinitò[1]

Abstract The present work shows a preliminary overview of the Big Data Analytics scenario, introducing the related privacy issues considered by the new General Data Protection Regulation, better known by its acronym GDPR. The work then introduces an innovative index to assess the compliance of a company with this regulation on the protection of personal data, in terms of privacy by design and privacy by default.

Abstract *Il presente lavoro mostra una preventiva panoramica in merito allo scenario del Big Data Analytics, introducendo le relative problematiche di privacy considerate dal nuovo Regolamento Generale sulla Protezione dei Dati, meglio noto con l'acronimo inglese GDPR. Il lavoro introduce quindi un innovativo indice per valutare la conformità di un'azienda a tale regolamento sulla tutela dei dati personali, in termini di privacy by design e privacy by default.*

Key words: GDPR, privacy, Big Data

1 Introduction

We live the historical moment of the Big Data boom, but even more often we realize that those who use this phrase do not really know its meaning: to understand well what Big Data is, we need to understand the deep meaning of the expression and how its influence can be noticed in everyday life. It is important to start by saying that the same phrase "Big Data" is somewhat misleading as it suggests the enormous amount of data available today in different sectors and, automatically, leads to the conclusion that Big Data revolution means opportunities to have so much information available

[1]Massimiliano Giacalone, Department of Economics and Statistics, University of Naples "Federico II"; massimiliano.giacalone@unina.it
Carlo Cusatelli, Ionian Department, University of Bari "Aldo Moro"; carlo.cusatelli@uniba.it
Fabio Fanari, iInformatica S.r.l.s., ffanari@iinformatica.it
Vito Santarcangelo, Department of Mathematics and Informatics, University of Catania, santarcangelo@dmi.unict.it
Diego Carmine Sinitò, iInformatica S.r.l.s., disinito@iinformatica.it

for business. This conclusion is only partially true, because there are sectors where data, although in large quantities, are not always available to everyone and, above all, are not always shared.

If Information Technology (IT) represents for Big Data the point from which to start with the necessary tools such as cloud computing, search algorithms, etc., on the other hand Big Data are necessary and useful in the most disparate business sectors as automotive, medicine, astronomy, biology, pharmaceutical chemistry, finance, gaming, commerce.

In the public sphere, there are many other types of Big Data applications:
- the deployment of police forces where and when the offenses are more likely to occur;
- the study of associations between air quality and health;
- genomic analysis to improve the resistance to drought of rice crops;
- the creation of models to analyze data coming from living beings in the biological sciences.

2 General Data Protection Regulation

As a consequence, the need arises to regulate the use of Big Data with the help of European legislation: the EU 679/2016 General Data Protection Regulation (GDPR) was born from this need, and the aim of this work is to provide an overview of the new legislation and to introduce a new index to measure GDPR compliance (Corrales M. et al, 2017). The GDPR, approved by the European Parliament in April 2016, will enter into force on May 25, 2018. The goal is to harmonize the laws on the confidentiality of information and privacy of all European countries and keep safe the sensitive user data processed by companies, and to limit uses according to the principles of (Anisetti et al., 2018):
- lawfulness, correctness and transparency: data must be processed in such ways;
- limitation of purposes: they must be determined, explicit and legitimate, then clearly identified;
- data minimization: data must be adequate, relevant and limited;
- accuracy: the data must be updated;
- restriction of storage: data must be kept for a limited period of time to achieve the purposes;
- integrity and confidentiality: adequate security of personal data must be guaranteed.

The GDPR, replacing the regulations of the individual European countries that differ from one another, represents an important step forward in terms of standardizing European policies and data protection at the continental level (Torra V., 2017). What changes is the extension of the jurisdiction to all companies that process personal data of subjects residing in the European Union, regardless of the geographical location of the company or the place where the data are managed and processed. Non-European companies that process data of European citizens will also have to appoint an EU

representative (Terry N., 2017). It is essential that European companies identify immediately how to adapt to the new legislation, thus avoiding being unprepared to face what is considered the most significant change in the history of data protection over the last 20 years.

It is necessary that companies immediately review their internal processes, placing user privacy as a primary element to guarantee priority and precedence. It is also necessary for companies to strengthen internal corporate communication through specific training programs so that anyone in a position that implies access to personal data of users correctly knows the extent to which they can carry out their profession. The concept of "privacy by design", a fundamental point on which the GDPR is concentrated (D'Acquisto, G., & Naldi, M., 2017) establishes that the data protection measures must be planned with the relative supporting IT applications starting from the planning of the business processes. This implies that only the data that are really indispensable for the performance of one's professional duties are processed and that access to information is limited only to those who have to carry out the processing. Another important point of the legislation concerns the "Breach Notification": data breach notifications are mandatory where the violation can put at risk the rights and freedoms of individuals. The notification must be made within 72 hours from the time the violation is verified and the customers are required to be informed "without undue delay".

The changes that the GDPR will bring are not only linked to the relationship between companies and users, but also concern the internal structure of the company: the new legislation will give greater prominence to the IT team and the company CIOs, making their tasks, nevertheless many managers still consider the GDPR as a waste of money and time, not understanding the importance of data protection today. (Mittal S. & Sharma P., 2017). With the GDPR, the figure of the Data Protection Officer (DPO) is established within the company with the task of monitoring the internal processes of the structure and acting as a consultant: the controllers of the monitoring and data processing activities are still required to notify their activities to local Data Protection Advisors (DPAs) which, for example within multinationals, can be a real bureaucratic nightmare, since each Member State has different notification requirements (Bertino, E., & Ferrari, E., 2018). With the introduction of the DPO, appointed on the basis of professional quality, expert in the field of law and data protection practices and equipped with the appropriate resources, the control of internal data management processes will be simplified.

The new legislation pays particular attention, in addition to what has already been said, to the requests for consent made to the subjects (Cohen M., 2017): the GDPR wants the requests to be submitted to the user in an "intelligible and easily accessible" manner, so that it is immediately clear what is the purpose of data processing. The companies will also have to guarantee users the right to delete personal data (Right to be forgotten), the possibility to request information about their treatment and to obtain a free copy in electronic format. The new regulation will be the cause of severe sanctions for companies that do not respect it, with fines of up to 4% of the total annual turnover or € 20 million, whichever is the greater of the two. But the

consequences will not be only economic: failure to comply with the new rules will also have repercussions on the reputation and image of the company, which will not be considered as attentive to the privacy of users and their sensitive data.

The GDPR has shed light on the issues of Data Protection (McDermott Y., 2017), a theme that, also due to the latest cyber attacks, requires ever more attention. It is well known that the threats against IT security and data protection are not going to decrease: just think of the recent attack of the WannaCry ransomware that hit more than 150 countries between Europe and Asia causing serious damage all over the world. Such a serious attack makes us understand the skills of today's hackers, always in search of flaws and inadequacies in IT systems, which must also be protected with the help of specialists in the sector. (Beckett P., 2017)

By taking advantage of effective security solutions, companies can protect themselves completely, thus guaranteeing their users that their data is always safe and that there is no risk of it being lost.

3 Innovative approach to GDPR compliance

To this end, it is essential to set up an IT infrastructure capable of analyzing the corporate GDPR compliance by analyzing in real time the various factors that feed the system. In particular, the system must be able to consider whether the activities of privacy by design and privacy by default are actually respected in the company. By privacy by design we mean that the company in planning a new service will have to ask itself if in this new service personal data will be processed, and if these data are ordinary or particular (sensitive): in this second case it will be necessary to express how these data will be protected. It is therefore necessary, from planning, to make all assessments concerning personal data if they are processed, the so called Data Protection Impact Assessment (DPIA).

For particular data we mean:

- patrimonial data, those related to income tax returns and other taxes and duties, etc.;
- any personal data that could potentially harm the dignity of the person or affect his natural right to privacy without legal reason.

On the other hand, the principle of privacy by default establishes that by default companies should only process personal data to the extent necessary and sufficient for the intended purposes and for the period strictly necessary for such purposes. It is therefore necessary to design the data processing system, ensuring that the collected data are not excessive. However, the UE 2016/679 does not present quantitative metrics to implement compliance about privacy by design and privacy by default. Then, it is very difficult to evaluate objectively if there is compliance on an infrastructure, and it lends itself to a heuristic implementation. Our approach introduces a possible methodology to evaluate the conformance about GDPR considering a metric to calculate privacy by design and by default considering the IT infrastructure of the company.

The risk level related to the GDPR of a business can therefore be defined as the following index:

GDPR_risk = risk_by_design + risk_by_default

If for convenience we indicate the two aforementioned addends with *rbds* and *rbdf*,

rbds = n_new/n_DPIA

where: *n_new* represents the number of services / products / processes that treat personal data activated by the business under study, *n_DPIA* represents the number of impact assessments on data protection.

A value close to 1 of *rbds* shows a high risk by design value.

Considering i the i-th asset, we can define rbdf with the following approach:

$$rbdf = \sum\left(\frac{c(i)}{BC(i) + DR(i)}\right) + VASS + PTEST$$

where the risk by default is given by various contributions: the first is related to asset management and is given by the sum of the impact of the individual assets (i), considering the ratio between the complexity (c_i) of the asset (i) and the value given by the presence of Business Continuity (BC_i) and Disaster Recovery (DR_i) for the considered asset (i), each contributes a value of 1; the incidence of the single asset (i) is thus mapped: in the case of low complexity such as desktop operating computers without specific data, the incidence of the asset is 0, in the case of medium complexity (data storage server) is equal to 1, in the case of high complexity (server for economic transactions) it is 2; the value of vulnerability assessment (*VASS*) is given by 0 in the case of established security measures (e.g. IPS, IDS), and penetration testing (*PTEST*) is dimensioned on the scale from 1 to 10 according to the Common Vulnerability Scoring System (CVSS).

A value of GDPR_risk below 10 is low (high compliance), 10 to 50 is medium (intermediate compliance), over 50 is high (no compliance).

This index is the basis of the GDPR_COMPLIANCE software developed by the young Sicilian company IINFORMATICA S.R.L.S. (first and unique innovative SME company of Trapani), free of charge for educational and academic purposes, that is a very useful tool to evaluate the GDPR compliance of one's system.

The following example represents the calculation of the *GDPR_risk* in an Italian SME with 2 locations (P1, P2) related to administration (which do not manage data locally on the machine), 1 data server (S1), 1 accounting server (S2) and 1 server for Disaster Recovery (S3). There is no backup of the 2 stations, but 1 backup of the data server and the accounting server are provided. 5 personal data treatment processes/services have been introduced with implementation of 5 DPIA. The data server and the accounting server process business data, but also special data (e.g. payroll), therefore fall within the average semantic category (average value equal to 1). There is an IPS device which then performs activities to guarantee a good Vulnerability Assessment (risk value 0), and a CVSS score equal to 1 has been found.

The value of the *risk_by_design* is given by 5/5 = 1.

The value of the *risk_by_default* is given by the following calculation:

$$rbdf = r(P1) + r(P2) + r\left(\frac{S_1}{S_3}\right) + r\left(\frac{S_2}{S_3}\right) + VASS + PTEST$$

$$rbdf = 0 + 0 + \left(\frac{1}{1}\right) + \left(\frac{1}{1}\right) + 0 + 1 = 3$$

GDPR_risk = *risk_by_design* + *risk_by_default* = 4.

The company has a low level of risk.

This index would be further reduced in the case of adoption of ISO 27001:2013 (information security management system) and ISO 22301:2012 (business continuity).

4 Conclusions

As far as data are really in unspeakable amounts, the real revolution referred to Big Data is the ability to use all this information to process, analyze and find objective evidence on different themes: the Big Data revolution refers precisely to what can be done with this amount of information, that is, to the algorithms capable of dealing with so many variables in a short time and with few computational resources. Until recently, to analyze a mountain of data that today we would call Small or Medium Data, a scientist would have taken a long time and would have used extremely expensive mainframe computers. Today, with a simple algorithm, the same information can be processed within a few hours, perhaps using a simple laptop to access the analysis platform. This presupposes new capacities to connect information to each other to provide a visual approach to data, suggesting patterns and models of interpretation so far unimaginable.

References

1. Anisetti, M., Ardagna, C., Bellandi, V., Cremonini, M., Frati, F., & Damiani, E. (2018). Privacy-aware Big Data Analytics as a Service for Public Health Policies in Smart Cities. Sustainable Cities and Society.
2. Beckett, P. (2017). GDPR compliance: your tech department's next big opportunity. Computer Fraud & Security, 2017(5), 9-13.
3. Bertino, E., & Ferrari, E. (2018). Big Data Security and Privacy. In A Comprehensive Guide Through the Italian Database Research Over the Last 25 Years (pp. 425-439). Springer, Cham.
4. Cohen, M. (2017). Fake news and manipulated data, the new GDPR, and the future of information. Business Information Review, 34(2), 81-85.
5. Corrales, M., Fenwick, M., & Forgó, N. (2017). New Technology, Big Data and the Law. Springer.
6. D'Acquisto, G., & Naldi, M. (2017). Big Data e Privacy by design (Vol. 5). G Giappichelli Editore.
7. McDermott, Y. (2017). Conceptualising the right to data protection in an era of Big Data. Big Data & Society, 4(1).
8. Mittal, S., & Sharma, P. (2017). General Data Protection Regulation (GDPR). Asian Journal of Computer Science And Information Technology, 7(4).
9. Terry, N. (2017). Existential challenges for healthcare data protection in the United States. Ethics, Medicine and Public Health, 3(1), 19-27.
10. Torra, V. (2017). Data Privacy: Foundations, New Developments and the Big Data Challenge. Springer International Publishing.