Chapter 4
Anonymity: A Comparison Between the Legal and Computer Science Perspectives

Sergio Mascetti, Anna Monreale, Annarita Ricci, and Andrea Gerino

4.1 Introduction

There are two opposing interests in our society: on one side, there is the need to collect and share information, which are activities that enable a number of services aimed at economic profit, scientific research, etc. On the other side, the right to personal data protection, intended as the right of disposal over all data in connection with our personality, requires to safeguard the subjects whose information is collected and shared. This contrast is one fragment of a broader problem concerning the relationship between law and technology. The overall question is whether legal definitions should adapt to technical solutions or if, vice versa, technology should implement the regulations in force. Certainly, the technological developments in the Internet era pose new questions to researchers in the two communities involved: Law and Computer Science. In this view, the topic of this paper, i.e., anonymity as a tool to guarantee personal data protection, is emblematic of the need for constant exchange of ideas and collaboration between these two communities.

The problem is that, despite the great research effort of both communities in the privacy protection field, most of the contributions address the problem either from the legal or the technical point of view only. This attitude has led to the

S. Mascetti (✉) • A. Gerino
Dipartimento di Informatica, University of Milan, Milan, Italy
e-mail: sergio.mascetti@di.unimi.it; andrea.gerino@di.unimi.it

A. Monreale
Department of Computer Science, University of Pisa, Pisa, Italy

Dipartimento di Informatica, University of Pisa, 3, Largo Pontecorvo, 56127 Pisa, Italy
e-mail: amonreale@di.unipi.it

A. Ricci
Department of Juridical Sciences "A. Cicu", University of Bologna, Bologna, Italy
e-mail: annarita.ricci@unibo.it

S. Gutwirth et al. (eds.), *European Data Protection: Coming of Age*,
DOI 10.1007/978-94-007-5170-5_4,

specification of basic definitions and objectives that only partially overlap, hence raising difficulties in communication and in the reciprocal applicability of the research results.

In contrast with this tendency, Ohm discusses the legal definitions of privacy, starting from the analysis of the contributions in the Computer Science community.[1] The conclusion presented in this paper is surprising: privacy law should not rely on the concept of anonymity. Jane Yakowitz's study[2] also leads to surprising conclusions. This paper addresses the problem of data anonymization for research purposes and it concludes that, since current privacy policies overtax valuable research without reducing any realistic risks, law should provide a safe harbour for the dissemination of research data and technical solutions are not necessary. In a recent paper, Schwartz et al.[3] support the idea that the concept of anonymity should be part of privacy laws, but its definition should be "reconceptualized". In these three papers, the interest resides, from our point of view, in their interdisciplinary approach.

With the aim to continue in the same direction, in this paper we attempt to integrate research on personal data protection in the two areas of Computer Science and Law. The approach is to address the central concept of anonymity from both perspectives, by reciprocally explaining the most important concepts, finding correspondences in the terminology and highlighting points in common and differences in the two areas. To achieve this, we first analyze the legal definitions of anonymous datum, as specified in the European Directive (Sect. 4.2). Then, we describe the main models and techniques proposed in the Computer Science literature to target the problem of anonymity (Sect. 4.3). Since this description of the state of the art in the two areas is targeted to readers in both communities, it focuses more on the main concepts and results, rather than on the technical details. We then discuss one similarity and some differences between the assumptions and definitions adopted by the two communities and the consequential results (Sect. 4.4). In particular, we focus on four main topics:

1. the role of anonymity in privacy preservation,
2. the relationship between identifying information and personal data,
3. the measurement of anonymity,
4. the relationship between anonymity and the principle of minimization.

We conclude that, despite there being some analogies, there are also a number of gaps, that on one side render some of the technical solutions not directly applicable to the regulations in force and, on the other side, suggest some specific interpretations

[1] Paul Ohm, "Broken Promises of Privacy: Responding to the Surprising Failure of Anonymization," *UCLA Law Review, Vol. 57, p. 1701, 2010* (2009).

[2] Jane Yakowitz, "Tragedy of the Data Commons," *Harvard Journal of Law and Technology, Vol. 25, 1, 2011*.

[3] Paul M. Schwartz and Daniel J. Solove, "The PII Problem: Privacy and a New Concept of Personally Identifiable Information," *New York University Law Review, Vol. 86, 2011* (2011).

of the current regulations in order to make them adequate to the existing technical solutions. Rather than a point of arrival, these conclusions are meant to be a starting point for discussion and integration between the two communities. In fact, thanks to its interdisciplinary character, this work tries to break down the communication barrier or at least the difficulties in dialogue between the two communities. The growing need of both communities for a systematic and interdisciplinary analysis of the anonymity notion and its use in protecting personal data can be adequately satisfied only through the development of a common language or at least a thorough understanding of the different approaches.

4.2 The Notion of Anonymity in European Legislation on Personal Data

The concept of anonymity has gained particular importance in relation to the application of European legislation on personal data. Indeed, while regulations apply to personal data, anonymous data are excluded from their field of application. This section analyses the legal understanding of anonymity, in particular with respect to the European Directive on personal data protection, and it tries to answer the following main questions:

- What is the interpretation of anonymity in common language?
- Should anonymity be considered a relative or absolute concept?
- What does anonymous data mean in legal terms?

To achieve this, we start with the notion of anonymity in common language (Sect. 4.2.1). Then we describe how the European legislation on personal data captures this concept.[4,5] Following the same approach of European legislation, we first introduce the concept of personal data (Sect. 4.2.2) and then proceed to defining anonymous data (Sect. 4.2.3). In order to show how European legislation has been implemented into national laws, we report the example of the anonymous data definition in the Italian Personal Protection Code (Sect. 4.2.4). The reason for choosing the Italian Personal Code is that it can be considered a "rigorous" implementation of the European Directive.

Before we proceed with the analysis, it is necessary to point out that, when we refer to the subjects of data processing, we use the definitions stated in Directive

[4] Directive 95/46/EC of the European Parliament and of the Council of 24 October 1995 on the protection of individuals with regard to the processing of personal data and on the free movement of such data, OJ L 281, 23.11.1995, 31–50.

[5] Directive 2002/58/EC of the European Parliament and of the Council of 12 July 2002 concerning the processing of personal data and the protection of privacy in the electronic communications sector (Directive on privacy and electronic communications), OJ L 201, 31.7.2002, 37–47.

95/46/EC: the *controller* is an entity (i.e., a natural or legal person, public authority, agency or any other body) that, alone or jointly with others, determines the purposes and means of personal data processing; the *processor* is an entity that processes personal data on behalf of the controller; the *recipient* is an entity to whom data are disclosed, whether a third party or not, and, finally, the *data subject* is the person to whom the personal data refer to.

4.2.1 The Notion of Anonymity in Common Language

In common language, the meaning of anonymity comes from the etymology of the term, that is, literally, "without name". "The word denotes an absolute concept: an anonymous person is one, of whom you do not know anything, somebody you cannot recognize or identify".[6] The definition of anonymity as an absolute concept is often taken for granted in the common understanding. However, as we will subsequently explain, anonymity in the legal context is actually a relative concept. Indeed, anonymity is often relative to specific facts, subjects and purposes. A musical arrangement, for instance, may be anonymous for a person but not for another, depending on whether this person knows the author. So the right to be anonymous, when recognized, refers to certain subjects, in predefined circumstances and for specific occasions, which can be specified by the law.[7] For example, the Italian legal system recognizes the biological mother's right not to be named in her son's birth certificate.

The transferral of the anonymity notion from common language to the legal context is not immediate. This is due to two main reasons. First, legal reasoning needs a degree of precision that is not generally required in common language. For instance, in legal terms it is necessary to specify the conditions that make a datum anonymous. Second, while the terms "anonymous" and "anonymity" are used in legal texts, they seem to have non-homogeneous values in the different legal sectors. In particular, we find references to the term "anonymous" in private law (copyright), criminal law (as an aggravating circumstance in some threat crimes), administrative law (open competitions for public recruitment) and constitutional law (freedom of expression). Consequently, we can conclude that the term "anonymity" is used in various areas but with a different slant, which makes it hard to extract a single univocal legal concept.

[6] Giusella Finocchiaro and Claire Vishik, "Law and Technology: Anonymity and Right to Anonymity in a Connected World," in *Movement-Aware Applications for Sustainable Mobility: Technologies and Approaches*, ed. Monica Wachowicz (IGI Global, 2010), 140-156.

[7] Giusella Finocchiaro, "Anonymity and the law in Italy," in *Lessons from the identity trail*, ed. Ian Kerr, Valerie M. Steeves and Carole Lucock (Oxford University Press, 2009), 523–536.

4.2.2 The Definition of Personal Data

The term "personal data" is defined as follows by Directive 95/46/EC:

> Personal data shall mean any information relating to an identified or identifiable natural person ("data subject"); an identifiable person is one who can be identified, directly or indirectly, in particular by reference to an identification number or to one or more factors specific to his physical, physiological, mental, economic, cultural or social identity.[8]

In the following we focus on three closely interrelated key elements of this definition:

1. "any information";
2. "relating to";
3. "an identified or identifiable".

1. The expression "any information" provides an idea of how wide the notion of personal data is. It is not infrequent to erroneously conceive "personal data" only as information concerning the most intimate aspects of a person. On the contrary, the concept of personal data includes any sort of information about a person, including economic and professional data, and not just data about his/her personal life. Indeed, this expression covers "objective" information, such as job or income as well as "subjective" information, such as opinions or assessments. This concept is also supported by Opinion 4/2007 of Article 29 Data Protection Working Party[9]:

   > Considering the format or the medium on which that information is contained, the concept of personal data includes information available in whatever form, be it alphabetical, numerical, graphical, photographical or acoustic, for example. It includes information kept on paper, as well as information stored in a computer memory by means of binary code, or on a videotape, for instance. In particular, sound and image data qualify as personal data from this point of view, insofar as they may represent information on an individual.

2. In general terms, information can be considered to "relate" to an individual when it is about that individual. In many situations, this relationship can be easily established. For instance, the data registered in a medical record are clearly "related to" an identified patient. Analogously, the image of a person filmed on a video interview is "related to" that person.

   In other situations, however, establishing the relationship between the information and the individual does not come immediately. In order to clarify this point, Article 29 Working Party noted that, "data relates to an individual if it refers to the identity, characteristics or behaviour of an individual or if such information is used to determine or influence the way in which that person is treated or evaluated".[10]

---

[8] Directive 95/46/EC, Art. 2.

[9] Opinion 4/2007 of Article 29 Data Protection Working Party on the concept of personal data, WP 136, 20.06.2007.

[10] Working Party document on data protection issues related to RFID technology, WP 105, 19/01/2005, Art. 8.

3. In general terms, a natural person can be considered "identified" when, within a group of people, he or she is "distinguished" from all other members of the group. Accordingly, the natural person is "identifiable" when, although the person has not yet been identified, it is possible to do so. This means that the subject can be identified through some characteristics or aggregation of data.

   Identification is normally based on particular pieces of information that we may call "identifiers" and which hold a close relationship with the given individual. Examples are outward signs of this person's appearance like height, eye colour, clothing, or a quality of the person that cannot be immediately noticed, like the profession, or the name. We will focus our attention on identifiers in Sect. 4.3.

4.2.3 The Concept of Anonymous Data

The concept of "anonymous data" is not explicitly reported in Directive 96/46/EC. However, this notion can be derived from the definition of "personal data" given in the Directive, and from some Recitals[11] of the same Directive. In particular, Recital no. 26 states that:

> The principle of protection must apply to any information concerning an identified or identifiable individual.

Furthermore:

> […] the principles of protection shall not apply to data rendered anonymous in such a way that the data subject is no longer identifiable.

Further references to "anonymous data" and especially to "anonymization" have been provided in Recitals no. 9, 26, 28 and 33 of Directive 2002/58/EC. In particular, Recital no. 9 states:

> The Member States (…) should cooperate in introducing and developing the relevant technologies where this is necessary to apply the guarantees provided for by this Directive and taking particular account of the objectives of minimising the processing of personal data and of using anonymous or pseudonymous data where possible.

Similarly, Recital no. 30 states that:

> Systems for the provision of electronic communications networks and services should be designed to limit the amount of personal data necessary to a strict minimum (…).

The above Recitals basically state the same principle in different ways: the principle of minimization in data processing. According to this principle, the processing of personal data is permitted only if it is required to achieve a specified purpose: if this very purpose can be accomplished with anonymous or pseudonymous data,

[11] The Recitals are the opening statements that introduce the main provisions of the European Directives and present the reasons for their adoption.

then these latter modalities should be preferred. Given these considerations, we can assume that in Directive 95/46/EC anonymity is considered as the main form of protection of the rights of the subjects whose data are processed.

4.2.4 *A Case Study: The Definition of Anonymous Data in the Italian Personal Protection Code*

Unlike Directive 95/46/EC, the Italian Personal Protection Code (or shortly "the Privacy Code") explicitly defines anonymous data as:

> (…) any data that, in origin or after being processed, cannot be connected to an identified or identifiable person.[12]

The Privacy Code definition has three key elements: the notion of data, the connection between the data and the person, and the identifiability of the latter one. These elements reflect the essential components of the definition of personal data comprised in Directive 95/46/EC.

The data. Briefly, we can assume that the definition of personal data in the Privacy Code, similarly to Directive 95/46/EC, is broad and it includes all information directly or indirectly related to a natural person.[13]

The connection. Both the Privacy Code and Directive 95/46/EC report that an essential element in the definition of anonymous data is the absence of a clear connection between the data and an identified (or identifiable) person. In fact, the distinction between anonymous and personal data actually depends on this connection. One problem is that, according to the definition of personal data given by the Privacy Code, all possible links between a person and information can be considered as personal data, and more subjects can be involved with multiple connections, as shown in Example 1.

Example 1 Consider a report made by a consultant Alice for a banker Bob concerning the financial situation of a client Carl applying for a loan. Alice is author of the report, and this fact is a personal datum related to Alice. Bob is the addressee of the report, and the fact that such a report is addressed to Bob is a personal datum related to Bob. Carl is the person having that financial situation, and the fact that such report concerns his very situation is a personal datum related to Carl. So, here we have three different data subjects, whose connections with personal data can be broken as to create three anonymous data.

---

[12] Italian Personal Protection Code, Legislative Decree no. 196, 30/06/2003, art. 4, co. 1, lett. n).

[13] A recent decision of the Italian Supreme Court (no. 19365, 22/09/2011) has stated the following principle: data about the health of a child is "sensitive data" (according to the definition of Legislative Decree no. 196/2003, art. 4, co. 1, lett. d) of the child's parents: therefore an unlawful processing of this information allows the parents to act for the protection of an own right.

Usually, unlike Example 1, a large amount of data is involved, and the relationship among the entities can be more complex. This example alone, however, highlights that anonymity is a relative and functional concept. In this example, in fact, anonymity would effectively be guaranteed by eliminating the connections between all the three parties involved in the report.

Identifiability. Which criteria should be followed to determine if a subject is identifiable? In Italy, as in other Member States, the evaluation of the measures of identification is carried out accordingly to European legal acts. In particular, Recommendation of the Council of Europe No. R (97) 5[14] specifies whether the impossibility of the connection between information and a person should be absolute or relative. This act states that information cannot be considered identifiable if identification requires an unreasonable amount of time and manpower.

A more accurate investigation of this matter can be found in the Explanatory Memorandum to Recommendation R (97) 18,[15] concerning the protection of personal data collected and processed for statistical purposes. See for instance point No. 52, letter d:

> Conditions for anonymity are relative, especially in relation to the technical means available for identifying data and taking away their anonymity. In this way, in view of the rapid progress in technological and methodological developments, the time and manpower required to identify a person, which would today be considered 'unreasonable', might no longer be so in the future (…).

Example 2 Data concerning "a graduated male living in Milan" would not be considered personal data, since it cannot be linked to a specific person, even if a great amount of time and manpower is used. Vice versa, data referring to "Sergio Mascetti, assistant professor at the University of Milan" should certainly be considered as personal data, since the identification of the person is immediate even with negligible time and manpower. However, it would not be as immediate to evaluate whether data referring to "a graduated male, living in Milan and working for a university, who plays volleyball and is a fan of Bruce Springsteen" should be considered personal data. What is hard to evaluate is how many persons correspond to this description and, even if there is a single one, it is not so clear as to how much time and manpower is required to identify him.

In order to address problems like the one reported in Example 2, it is necessary to analyze each case in its different aspects, taking into account all the following factors, as stated by Opinion 4/2007: the intended purpose of data processing, the way the processing is structured, the advantage expected by the controller, the interests at stake for the individuals, and the risk of organisational dysfunctions and technical

[14] Recommendation No. R (97) 5 of the Committee of Ministers to Member States on the protection of medical data, 13/02/1997.

[15] Recommendation No. R (97) 18 of the Committee of Ministers to Member States on the protection of personal data collected and processed for statistical purposes, 30/09/1997.

failures. The identification process is dynamic and "should consider the state of the art in technology at the time of the processing and the possibilities for development during the period for which the data will be processed".[16]

Observe that in the acts mentioned above the concept of reasonableness is used to assess identifiability. This concept is commonly used in legal systems as a measurement criterion. In this perspective, reasonableness is the criterion used to measure how "easy" it could be to associate a data subject with the data. This approach remarks the fact that anonymity is a relative concept, and its evaluation requires taking into account the particular context at the time of processing.

The degree of anonymity cannot be predetermined: in fact, anonymity may take a different extent depending on the circumstances, among which we may include the will of the data subject. It is therefore essential to suggest some criteria for measuring anonymity. The possible quantification of anonymity will be analyzed from a technological point of view in the next section.

4.3 Anonymity in Data Disclosure

In this section we briefly survey some of the contributions in the Computer Science literature for the problem of guaranteeing anonymity while disclosing data. Note that we have decided to focus our discussion on anonymity models, thereby omitting many other interesting models, such as randomization[17] and differential privacy,[18] whose purpose is to alter the private information, rather than render a data respondent anonymous.

We consider two of the applicative scenarios that have been mainly addressed by the research community: data publication (Sect. 4.3.1) and location based services (Sect. 4.3.2).

4.3.1 Anonymity in Data Publication

As we observed in Sect. 4.2, the disclosure of personal information to the general public or to third parties is subject to the limitations imposed by the regulations on privacy protection. Nevertheless, if this information was rendered anonymous, these

[16] Opinion 4/2007, Art. 12.

[17] Rakesh Agrawal and Ramakrishnan Srikant, "Privacy-preserving data mining," in *Proceedings of the 2000 ACM SIGMOD international conference on Management of data* (New York, NY, USA: ACM, 2000), 439-450.

[18] Cynthia Dwork, "Differential Privacy," in *Automata, Languages and Programming,* 4052:1-12, Springer Berlin/Heidelberg, 2006.

Table 4.1 Hospital database

Name	Gender	Date of birth	ZIP code	Disease
Alice	F	01/01/1981	11111	Flu
Anne	F	02/02/1981	11122	Flu
Sonia	F	12/03/1981	11133	Flu
Bob	M	12/01/1982	33311	Heart disease
Shunsuke	M	10/04/1982	33322	Cold
Carl	M	02/03/1982	33333	Flu

limitations would not apply, hence making it possible to share the information without explicit user agreement and with great benefits both for the entity collecting this information and the other stakeholders. For this applicative reason, the problem of rendering information anonymous before publication has been extensively studied in the scientific literature.[19] In this section we first describe the problem in detail and then survey some of the contributions addressing this problem.

4.3.1.1 Problem Definition and Characterization

The actors involved in a typical data publication scenario are the same described in Sect. 4.2, with the only difference that the controller and the processor are considered as a single entity; for this reason, in the following, when we mention the "controller" we refer to both the controller and the processor. The data flow is the following: the controller collects data from the subjects and wants to release this information to a recipient that can be, for example, a data miner or an analyst. Since we consider that the controller is trusted[20] by the data subject, the overall privacy problem is the following: guaranteeing the data subject's privacy protection, while releasing useful information to the recipient that plays the role of the *adversary*.

Example 1 Consider a hospital (i.e., the data collector) in which patient information (e.g., diseases, therapies, etc.) is collected and stored. Table 4.1 shows an example of this information.

This data is potentially a valuable resource for medical research (i.e., the recipient), but it cannot be disclosed without the user's explicit authorization, due to the regulation in force hence it needs to be altered before disclosure. Figure 4.1 shows a graphical representation of this situation.

[19] Anna Monreale, Dino Pedreschi, and Ruggero G. Pensa, "Anonymity technologies for privacy-preserving data publishing and mining," in *Privacy-Aware Knowledge Discovery: Novel Applications and New Techniques*, F. Bonchi, E. Ferrari, Chapman & Hall/CRC Data Mining and Knowledge Discovery Series, 2010.

[20] Here, the term "trust" is not used here in its proper legal sense but according to its intuitive meaning of "confidence". In this case, it means that the data subject is confident that the data collector will manage his/her data according to the current regulations or to other agreements between the two parties.

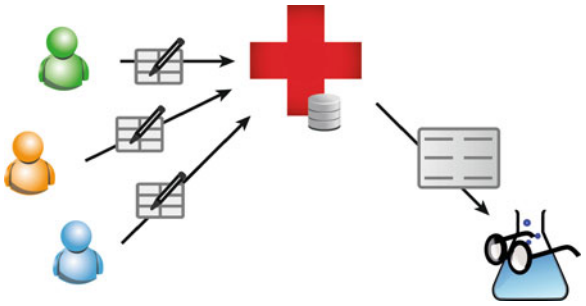

Fig. 4.1 Data flow in the data publication scenario

The problem with protecting data subject privacy when disclosing information is not trivial. Among many others, one intuitive reason is the following: providing data utility and data subject's privacy are contrasting objectives.[21] Indeed, a naïve solution to achieving the best data utility is to provide the recipient with exactly the same information collected by the controller. However, in this case the data subject's privacy is compromised. Vice versa, the best privacy protection is achieved when no data are disclosed, but in this case data utility is null. This is one of the reasons that make the problem scientifically attractive and that have led it to be extensively studied by the Computer Science and the Official Statistics communities. Both communities proposed several mathematical representations of the problem, considering different aspects of it. These mathematical representations, that we call *privacy models*, have two main objectives: to formally describe the problem and to make the correctness of the privacy preserving techniques possible to prove.

Each privacy model defines all the important aspects of the considered problem, like the actors, the flow of data (i.e., collection and successive release), etc. In particular, most of the privacy models defined in the literature identify one aspect that is particularly important: the *attack model*. With this term we indicate the adversary's capabilities used in his attempt to discover the data subject's personal information. These capabilities include the *inference abilities* (i.e., how to derive new information from the existing one) and, in particular, *the background knowledge*, i.e., the information that the recipient owns independently from the data released by the controller. Background knowledge can be originated by several sources, such as well-known facts, demographic information, public records, and information on specific individuals possibly published by the data subject himself (e.g., data published in a social network).

In order to continue with this discussion, it is necessary to better characterize the type of information collected by the controller. Many of the contributions identify four groups of attributes[22] (e.g., each column in Table 4.1 is an attribute):

[21] Tiancheng Li and Ninghui Li, "On the tradeoff between privacy and utility in data publishing," in *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining* (New York, NY, USA: ACM, 2009), 517-526

[22] Valentina Ciriani et al., "Microdata Protection," in *Secure Data Management in Decentralized Systems*, Springer US, 2007, 33:291-321.

- Explicit identifiers of the data subject, such as name and social security number.
- Quasi-Identifiers (QI): attributes that are not explicit identifiers but that, when used in conjunction with background knowledge, can lead the adversary to identify a data subject or to restrict the possible identity of a data subject; the attributes "gender", "ZIP code" and "date of birth" are examples of QI.
- Private Information (PI): personal data that should not be associated to a data subject's identity like, for example, a disease or salary.
- Non-private information: all the attributes that do not fall into the previous categories.

4.3.1.2 *k*-Anonymity

Samarati et al.[23] showed that simply dropping the explicit identifiers does not guarantee anonymity if the adversary knows the population's QI values (this information can be obtained, for example, from the voter list). In this case, referring to Example 1, the adversary can discover that there is a single male person born on the 12/01/1982 who lives at ZIP code 33311. Since this information in the voter list is associated to an explicit identifier (i.e., the name), the adversary can discover that Bob had the flu. This type of attack is sometimes called *record linkage attack.*[24] Typically, a countermeasure against this attack is to apply a transformation to the values in the QI attributes in order to render several records indistinguishable.

A well-known model, defined to contrast the record linkage attack, is *k*-anonymity.[25] This approach became popular in the field of privacy preserving data publication and in many other privacy problems. The idea of *k*-anonymity is to guarantee that information on any data subject cannot be distinguished from the information on other k-1 data subjects. More technically, the privacy requirement defined by *k*-anonymity is that for each record released (e.g., a record is a row in a table) there must be at least other $k-1$ records with the same QI values. The techniques adopted in the literature to enforce *k*-anonymity involve the removal of explicit identifiers and the generalization (e.g., the date of birth is replaced by the year of birth) or suppression (e.g., removing the date of birth) of QI. It is evident that these techniques reduce the accuracy of the disclosed information.

[23] Pierangela Samarati and Latanya Sweeney, "Generalizing data to provide anonymity when disclosing information (abstract)," in *Proceedings of the seventeenth ACM SIGACT-SIGMOD-SIGART symposium on Principles of database systems, PODS '98* (New York, NY, USA: ACM, 1998).

[24] William E. Winkler, *The state of record linkage and current research problems* (Statistical Research Division, U.S. Bureau of the Census, 1999), Washington, DC.

[25] Id. at 17. ("Generalizing data to provide anonymity when disclosing information (abstract)").

Table 4.2 A 3-anonymous version of Table 4.1.

QI attributes			PI attribute
Gender	Date of birth	ZIP code	Disease
F	1981	111*	Flu
F	1981	111*	Flu
F	1981	111*	Flu
M	1982	333*	Heart disease
M	1982	333*	Cold
M	1982	333*	Flu

*denotes that some information has been removed to guarantee anonymity.

Table 4.3 A 3-anonymous table with respect to quasi identifiers QI_1 and QI_2

Gender	Date of birth	ZIP code	Disease
F	1981	333*	Flu
F	1982	111*	Flu
F	1982	111*	Cold
M	1982	111*	Heart disease
M	1981	333*	Cold
M	1981	333*	Flu

*denotes that some information has been removed to guarantee anonymity.

Example 2 Table 4.2 represents a 3-anonymous version of Table 4.1. Note that Table 4.2 reports the year of birth only (instead of the birthdate) and that the last digits of the ZIP Code have been suppressed. In this case, even if the adversary knows the Gender, Date of Birth and ZIP Code of the entire population, he would not be able to distinguish Bob's record from the records of other two users (Shunsuke and Carl).

4.3.1.3 *k*-Anonymity with Multiple QI

Models based on *k*-anonymity assume that the controller knows the QI. However, different adversaries may use different QIs. To address this problem, one extension to *k*-anonymity consists in making multiple QIs possible to specify.[26] In other words, the controller knows a set of quasi-identifiers and the disclosed information has to be k-anonymous with respect to each of them. Example 3 shows that guaranteeing k-anonymity for all the quasi-identifiers in a set Q is not the same as guaranteeing k-anonymity on a QI that is the "union" of all the quasi-identifiers composing Q.

Example 3 Consider the data represented in Table 4.3. Assume that the controller identifies two sets of QI: $QI_1 = \{Gender\}$ and $QI_2 = \{Date\ of\ Birth, ZIP\ Code\}$.

Table 4.3 is 3-anonymo with respect to QI_1 and QI_2, but it is not 3-anonymous when the quasi identifier is $QI_1 \cup QI_2$, i.e., $QI = \{Gender, Date\ of\ Birth, ZIP\ Code\}$. Indeed, there is one group of three records with Gender = "F" and another group of three records with Gender = "M". Similarly, considering QI_2, we can identify two

[26] Benjamin C. M. Fung, Ke Wang, and Philip S. Yu, "Anonymizing Classification Data for Privacy Preservation," *IEEE Trans. on Knowl. and Data Eng.* 19, no. 5 (May 2007): 711–725.

Table 4.4 A 3-anonymous database

Gender	Date of birth	ZIP code	Disease
F	1981	111*	Flu
F	1981	111*	Flu
F	1981	111*	Flu
M	1982	333*	Heart disease
M	1982	333*	Cold
M	1982	333*	Cold

*denotes that some information has been removed to guarantee anonymity.

different groups, each one with three indistinguishable records with respect to the Date of Birth and ZIP Code. However, the table is not 3-anonymous with respect to the set QI = {Gender, Date of Birth, ZIP Code}. For example, there is a single record with the combination Gender = "F", Date of Birth = "1981" and ZIP Code = "333*".

4.3.1.4 *l*-Diversity

The models illustrated in Sects. 4.3.1.2 and 4.3.1.3 aim to avoid that any record in a table can be associated with less than k individuals. However, this property is not sufficient to guarantee an intuitive notion of anonymity. Indeed, it has been shown that, although the adversary may not uniquely identify the data subject "referred" by a record, he can still infer the personal information of that individual. Two attacks have been presented in the literature to achieve this.[27] The former, called "homogeneity attack" is based on a vulnerability of the k-anonymity model and is intuitively explained in the following example.

Example 4 Consider Table 4.2. Suppose that the adversary knows that Alice was born in 1981, lives in the area with ZIP code 11111 and is in the database. He knows that Alice's record is one of the first three in the table. Since all of those patients have the same medical condition (Flu), the adversary can identify Alice's disease.

The latter attack that can be used to violate the data subject's privacy despite k-anonymity, is called "background knowledge attack" since it assumes that the adversary has additional background information. This attack is based on the idea that in some cases there can be a correlation between the QI values and the private information. Consider the following example.

Example 5 Consider the 3-anonymous Table 4.4 and suppose that the adversary knows that Shunsuke is in the database, was born in 1982 and is Japanese.

The attacker can infer that Shunsuke's record is one of the last three records in the above table. Also, by knowing that Japanese people have a low incidence of heart disease, the adversary can conclude with high likelihood that Shunsuke has a Cold.

[27] Ashwin Machanavajjhala et al., "l-diversity: Privacy beyond k-anonymity," *ACM Trans. Knowl. Discov. Data* 1, no. 1 (March 2007): 24.

Table 4.5 A database satisfying 3-diversity

Gender	Date of birth	ZIP code	Disease
F	1981	111*	Flu
F	1981	111*	Cancer
F	1981	111*	Cold
M	1982	333*	Heart disease
M	1982	333*	Flu
M	1982	333*	Cold

It is worthwhile observing that there is a significant conceptual difference between the two attacks above. The former (i.e.: the "homogeneity attack") takes place under the same assumptions specified for *k*-anonymity and exploits a vulnerability of this model. Vice versa, the latter (i.e.: the "background knowledge attack") exploits some background knowledge that the *k*-anonymity model assumes as not available to the attacker. Note that, in general, given a privacy preserving technique that is safe under a privacy model, it is always possible to find a counter example to show that that technique is insufficient (or "unsafe") by using more background knowledge than assumed in that privacy model.

The *l-diversity* model was proposed in order to overcome the weakness of *k*-anonymity and to counter the two attacks illustrated above.[28] The aim is to obtain groups of data subjects with indistinguishable QIs and an acceptable diversity of the attributes' values representing personal information. In particular, the main idea of this method is that every k-anonymous group should contain at least *l* values for the attributes containing personal information. Different instantiations of the *l*-diversity definition have been presented by Machanavajjhala et al.[29] and Xiao et al.[30]

Example 6 Consider the database represented in Table 4.5. It satisfies 3-diversity and it is safe against the attacks illustrated in Examples 4 and 5. Indeed, the adversary cannot understand if Alice suffers from "Flu", "Cancer" or "Cold". Moreover, when the adversary tries to identify Shunsuke's disease, after excluding "Heart Disease", there are still two other possible diseases.

4.3.1.5 *t*-Closeness

It has been observed that in some cases the *l*-diversity model can lead to unnecessary generalization, if we consider different degrees of "sensitivity" of private information. This is better explained by the following example.

[28] Id. at 21 ("l-diversity: privacy beyond k-anonymity").

[29] Id.

[30] Xiaokui Xiao and Yufei Tao, "Personalized privacy preservation," in *Proceedings of the 2006 ACM SIGMOD international conference on Management of data*, SIGMOD '06 (New York, NY, USA: ACM, 2006), 229–240.

Table 4.6 A k-anonymous database

Age	ZIP code	Disease
[21–30]	111*	Negative
[21–30]	111*	Negative
[21–30]	111*	Negative
[21–30]	111*	Negative
[41–45]	222*	Negative
[41–45]	222*	Positive
[41–45]	222*	Negative
[41–45]	222*	Positive
[31–40]	111*	Positive
[31–40]	111*	Positive
[31–40]	111*	Positive
[31–40]	111*	Negative
…		
[60–70]	444*	Negative

*denotes that some information has been removed to guarantee anonymity.

Example 7 Consider the data in Table 4.6 where the attribute "Disease" contains the value "Negative" for patients with a negative HIV test result and the value "Positive" for those with a positive test result. Assume that in this table we have 10,000 records and only 1% of them has Disease = "Positive". Clearly, the two values have a different degree of sensitivity. Intuitively, a patient with a negative test result would not mind the result being known, because it is the same as that of 99% of the population, but he/she would not want to disclose a positive value. Therefore, the level of anonymity required for the first group in Table 4.6 (i.e., age "[21–30]", ZIP code "111*") is intuitively weaker than the one required for the second group (age [41–45], ZIP code "222*").

Another problem with *l*-diversity is that it can be insufficient to prevent the disclosure of private information when the adversary knows the distribution of the private values. Indeed, if the adversary has prior knowledge about private information on a data subject, he can compare this knowledge with the probability computed from observing the disclosed information. In Example 7, the adversary knows that the average distribution of positive HIV persons is 1%. After observing the disclosed information, the adversary discovers that Bob (age 32 and living in ZIP code 11123) has a much higher probability to be HIV positive (i.e., 75%).

In order to avoid the above weakness of *l*-diversity, Li et al. introduced the *t*-closeness model.[31] This technique requires that in any group of QIs the distribution of the values of an attribute containing personal information is close to the distribution of the attribute values in the overall table. The distance between the two distributions should be no more than a threshold *t*. Clearly, this limits the information gained by the adversary after an attack.

[31] Ninghui Li, Tiancheng Li, and S. Venkatasubramanian, "*t*-closeness: Privacy Beyond k-Anonymity and l-Diversity," in *Data Engineering, 2007. ICDE 2007. IEEE 23rd International Conference on*, (Istanbul, Turkey: IEEE Computer Society, 2007) 106–115.

4.3.2 Anonymity When Disclosing Spatio-Temporal Information

So far, most of the techniques illustrated in this section assume that the data to disclose are either in the form of numbers (e.g., the age, the salary, etc.) or elements organized in taxonomy (e.g., gender, diseases, etc.). Several contributions investigate the problem of guaranteeing users' anonymity in presence of spatio-temporal information. We first describe the problem (Sect. 4.3.2.1) and then introduce the models and techniques proposed in the Computer Science literature to address it (Sect. 4.3.2.2).

4.3.2.1 Problem Description

Some preliminary contributions motivate that specialized techniques are required in presence of spatio-temporal information,[32,33] This is mainly due to three reasons. First, it is commonly recognized that this kind of information has a very specific semantic that calls for specialized data managements methods. Secondly, most of the techniques related to data publication (like the ones introduced in Sect. 4.3.1) assume that each data subject is associated with a fixed amount of information (e.g., a single record), while many of the applications that involve spatio-temporal information associate a list of locations (also called a "trace") with each user. The last, but conceptually most important reason, is that in many practical cases, space and time can have the double role of quasi identifiers and of private information (see Example 8 below).

Spatio-temporal information is particularly relevant from an applicative point of view, because it is the fundamental data type in geo-referenced applications and services that are becoming popular mainly thanks to the diffusion of mobile devices (e.g., smartphones). These devices are "location-aware" in the sense that they are equipped with hardware peripherals that make their geographical location possible to detect. This new feature gives raise to a new class of Internet services, called *Location Based Services* (LBS), in which one of the parameters of the requests is the current location of the user. One example of LBS is the "find the closest Point of Interest (POI)" where a POI is, for instance, a restaurant. In this context, privacy should be safeguarded both when each request is issued (this is sometimes called the "on-line" privacy protection problem) and when a dataset of formerly acquired location information needs to be disclosed (i.e., the "off-line" privacy protection problem).

The actors in this scenario are similar to the ones in the data publication scenario. In the "off-line" privacy protection problem the *user* (i.e., the data subject) reports

[32] Marco Gruteser and Dirk Grunwald, "Anonymous Usage of Location-Based Services Through Spatial and Temporal Cloaking," in *Proceedings of the 1st international conference on Mobile systems, applications and services*, MobiSys '03 (New York, NY, USA: ACM, 2003), 31–42.

[33] Sergio Mascetti et al., "k-Anonymity in Databases with Timestamped Data," *in Proceedings of the Thirteenth International Symposium on Temporal Representation and Reasoning* (Washington, DC, USA: IEEE Computer Society, 2006), 177–186.

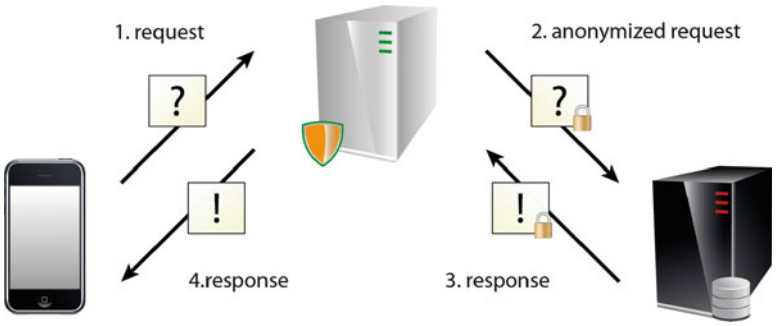

Fig. 4.2 Data flow in the provisioning of a LBS service with anonymization

Fig. 4.3 (**a**) *On the left*, reported users' locations. (**b**) *On the right*, identification of commonly visited places

his/her locations to a trusted *location server* (i.e., the controller and processor) that collects the information. After proper modifications, the location server discloses the location information to a third party (i.e., the recipient), which is not trusted by the user. On the contrary, in the "on-line" privacy protection problem the user communicates with the *service provider* that is not trusted by the user, thereby playing the role of the recipient. In this case, the role of controller and processor is played by a trusted entity, called *anonymizer*, which is in charge of enforcing the user's anonymity. As shown in Fig. 4.2, a user issues a LBS request to the anonymizer, that properly modifies and forwards it to the service provider. The anonymizer also forwards the reply from the service provider to the user.

Example 8 Let's consider an LBS in which an "anonymous" user frequently reports his/her location (see Fig. 4.3a). By observing this information, the service provider can identify two recurring places from which most of the requests are issued (see Fig. 4.3b). The temporal information indicates that the reports from one of these two places occur during working hours, while the ones from the other place occur during non-working hours. Given this analysis, the service provider can conclude,

Fig. 4.4 Example of location 3-anonymity

with high likelihood, that the two places are the user's home and work place. From public sources, like a phone book, the service provider can compute the set of people living in that home address and working in that workplace. If the intersection of these two sets contains one person, the adversary can re-identify the user. Moreover, from the analysis of the reported locations, the service provider can also observe that there are other usual places for that user. One of these places is a Church, from which the user generally reports locations on Sunday morning. Given this observation, the service provider can deduce, with high likelihood, the user's religious belief.

4.3.2.2 Privacy Models for LBS Anonymity

The core idea of the defence techniques based on anonymity is to alter each request so that the exact location is transformed into a "generalized region" in such a way that an adversary cannot identify the possible issuer in a set that contains at least k users (see Example 9).

Example 9 Consider Fig. 4.4. The position labelled "A" is the current location of Alice, who is issuing an LBS request. The other markers represent the location of other four users. Assume that the adversary's background knowledge includes the identities and the corresponding positions of all five persons. Even if Alice removes any of the explicit identifiers from the LBS request, the adversary can re-identify her if Alice's exact location is reported. Vice versa, if the location of Alice is generalized to the dark-grey rectangle represented in Fig. 4.4 before the request is sent to the service provider, the adversary cannot identify the issuer of the request in the set of three persons, hence guaranteeing a form of 3-anonymity to Alice.

It is important to observe that the attack illustrated in Example 9 requires the adversary to have background knowledge that associates each user's location with the identity of that user. One problem is modelling how much information the adversary has. Indeed, on one hand, there is a common agreement about the fact that an adversary can partially obtain this background knowledge like, for example, the

information exploited by the adversary in Example 9. On the other hand, the adversary is unlikely to have "full background location knowledge", i.e., to know the location of each person in each time instant. In other words, the adversary has "partial background location knowledge" and the problem is how to model it.

In order to tackle this problem, a common approach is to assume that the anonymizer ignores the background information available to the adversary. In this case, it is assumed that the adversary always has "full background location knowledge". This is a "conservative" approach in the sense that if a defence technique is proved as safe under this assumption, it can be proved as safe in any case of partial background knowledge,[34,35,36,37] The drawback of this approach is that, by assuming "full background location knowledge", the anonymizer needs to generate large generalized regions that may render the service impractical. Some papers tackle this problem by assuming that the anonymizer can estimate an upper bound for the background knowledge available to the adversary and this bound is less than the "full background location knowledge". The advantage of the techniques proposed under this assumption is that the generalized region, required to achieve anonymity, is generally smaller,[38,39] However, the problem with this approach is that if the assumption about the adversary knowledge is incorrect, and the adversary actually has more background knowledge than assumed, then there are no guarantees on the actual anonymity of the disclosed information.

The first paper addressing the problem of guaranteeing *k*-anonymity when providing an LBS service considers an adversary with "full background location knowledge".[40] Although on one side this model is conservative, it has been shown that, from other perspectives, this model is not sufficiently conservative, leading to possible privacy breaches. Two formal models independently proposed by Kalnis et al.[41] and Mascetti et al.[42] capture this problem. The intuition is the following: the attack

[34] Id. at 29 ("Anonymous Usage of Location-Based Services Through Spatial and Temporal Cloaking").

[35] Mohamed F. Mokbel, Chi-Yin Chow, and Walid G. Aref, "The new Casper: query processing for location services without compromising privacy," in *Proceedings of the 32nd international conference on Very large data bases*, VLDB '06 (Seoul, Korea: VLDB Endowment, 2006), 763–774.

[36] Panos Kalnis et al., "Preventing Location-Based Identity Inference in Anonymous Spatial Queries," *IEEE Trans. on Knowl. and Data Eng.* 19, no. 12 (December 2007): 1719–1733.

[37] Sergio Mascetti et al., "Spatial generalisation algorithms for LBS privacy preservation," *J. Locat. Based Serv. 1*, no. 3 (September 2007): 179–207.

[38] Claudio Bettini et al., "Anonymity in Location-Based Services: Towards a General Framework," in *Proceedings of the 2007 International Conference on Mobile Data Management* (Washington, DC, USA: IEEE Computer Society, 2007), 69–76.

[39] Manolis Terrovitis and Nikos Mamoulis, "Privacy Preservation in the Publication of Trajectories," in *Proceedings of the Ninth International Conference on Mobile Data Management* (Washington, DC, USA: IEEE Computer Society, 2008), 65–72.

[40] Id. at 29 ("Anonymous Usage of Location-Based Services Through Spatial and Temporal Cloaking").

[41] Id at 35 ("Anonymity in Location-Based Services: Towards a General Framework").

[42] Id. at 36 ("Privacy Preservation in the Publication of Trajectories").

Fig. 4.5 Failure of location anonymity in the "historical case"

model considered by Gruteser et al. implicitly assumes that the adversary does not know the defence technique. If this assumption does not hold, which is often the case, the defence technique proposed by Gruteser et al. may fail to provide the required level of anonymity.

Another limit of some existing models,[43,44,45,46] is assuming that the adversary cannot associate two or more requests with the same user. This assumption is sometimes called the "snapshot case" since it is equivalent to assuming that the adversary can observe the users' positions and requests in a given instant and cannot "follow" the users' movements. However, in many practical cases, each user is associated with a pseudo-id (a unique value, whose association with the real user identity is kept secret) that is sent by the user with each request. In this "historical case" the adversary can understand that a single user issues two or more requests. It has been shown that this knowledge may render ineffective the defence techniques proposed for the "snapshot case" (see Example 10). This problem has been addressed, among others, by Bettini et al.[47] and Riboni et al.[48]

Example 10 Consider Fig. 4.5 that represents the locations of five users in two different time instants. Alice is the user labelled "A" who issues two LBS requests, one in each time instant. According to the intuitive definition of *k*-anonymity provided above, the two dark-grey rectangles reported in the figure guarantee a form of

[43] Id. at 29 ("Anonymous Usage of Location-Based Services Through Spatial and Temporal Cloaking").

[44] Id. at 32 ("The new Casper: query processing for location services without compromising privacy").

[45] Id. at 35.

[46] Id. at 36.

[47] Claudio Bettini, "Privacy and anonymity in Location Data Management," in *Privacy-Aware Knowledge Discovery: Novel Applications and New Techniques*, ed. F. Bonchi, E. Ferrari, Chapman & Hall/CRC Data Mining and Knowledge Discovery Series, 2010.

[48] Daniele Riboni et al., "Preserving Anonymity of Recurrent Location-Based Queries," in *Proceedings of the 2009 16th International Symposium on Temporal Representation and Reasoning*, TIME '09 (Washington, DC, USA: IEEE Computer Society, 2009), 62–69.

3-anonymity. However, if the adversary is able to understand that a single user issued both requests, the only possible issuer is Alice since she is the only user that is located within both rectangles.

4.4 Discussion

In this section lawyers and computer scientists "talk to each other". After the analysis of the anonymity concept, conducted in accordance with traditional approaches in both areas, we now highlight the main similarities and differences between the Legal and Computer Science fields. We argue that a "neutral" study of the two approaches is necessary to obtain a complete picture of the problem. This result should then be used as a starting point for innovative research in the area of privacy protection. We do not assume that one or the other approach is wrong or entails unsolvable problems and that it should, consequently, be changed and adapted to the other. By putting aprioristic statements aside, we aim to analyze both approaches under the same perspective, which is based on a systematic examination of the problem, starting with a detailed linguistic and formal analysis.

4.4.1 The Role of Anonymity in Privacy Preservation

As observed in Sect. 4.2, the legal notion of anonymity, as defined in the legislation on data protection, cannot be seen as a right in itself. Instead, anonymity should be considered as a "tool" that can be used to safeguard the protection of personal data. This interpretation is compatible with the current approach adopted in Computer Science. Indeed, although most of the scientific contributions tackle the problem of guaranteeing privacy through anonymity, it has also been recognized that privacy protection can also be achieved without anonymity. Consider the following example.

Example 11 Assume a geo-referenced social network in which each user can share his/her location with some friends. Note that, if we address the privacy problem of a user Alice with respect to her friend Bob (i.e., Bob is the adversary), anonymity cannot be used to protect privacy, since the service requires Bob to know which user is located in a given location. Also, pseudonyms are not effective, since in many cases Bob knows Alice in person. One solution that Alice can adopt to protect her privacy is to avoid using the service or to exclude Bob from the list of users enabled to see her location. However, the question is whether it is possible to allow Alice and Bob to enjoy the service, while still providing a form of privacy protection. One solution is to allow Alice to specify her "privacy preference" in terms of an "obfuscated area": Bob will only be able to understand that Alice is in that area, and the adopted technique ensures that Bob cannot understand where Alice is located within

Fig. 4.6 Two examples of "obfuscated area"

that area. Figure 4.6 shows Alice's actual position (that is hidden from Bob) and two possible "obfuscated areas" (the dark-grey rectangles), the larger one providing a higher level of privacy protection.

As shown in Example 11, when it is not possible or convenient to render the data anonymous, one approach is to allow each user to specify which information is "sensitive" (accordingly to the will of the data subject) and to guarantee that only "non-sensitive" information is disclosed. Determining whether these techniques are supported by sound legal bases is out of the scope of this paper, but it certainly is an interesting research topic. Indeed, from a legal point of view, the problem cannot be easily solved. The law's requirement, in a general sense, is to protect the fundamental rights of the individuals, giving equal importance to all information, without any difference in value. In particular, the issues concerning the possibility of allowing each data subject to choose the preferred level of privacy have still not been extensively addressed in European directives.

4.4.2 Identifying Information and Personal Data

Another point in common between Law and Computer Science is that both recognize the relative nature of anonymity. In particular, the intuition that simply dropping explicit identifiers is not sufficient to guarantee anonymity is formulated in the legal context (e.g., see Sect. 4.2.3) and it is also supported by formal models presented in the scientific literature (among the others, in Samarati et al.[49] and Gruteser et al.[50]). Indeed, although legal norms do not explicitly distinguish between "explicit identifiers" and "quasi identifiers", this distinction is compatible with the current legal approach.

[49] Id. at 17 ("Generalizing data to provide anonymity when disclosing information (abstract)").

[50] Id. at 29 ("Anonymous Usage of Location-Based Services Through Spatial and Temporal Cloaking").

Vice versa, the specification of "personal" (or "private") information is different in the two areas of Law and Computer Science. Indeed, as explained in Sect. 4.2.2, the term "personal data" denotes any kind of information about a person, including information that is intuitively "sensitive" (like religious beliefs) and those that are not (like eyes colour). Also, the term "personal data", as intended in legal norms, refers both to information that should only be known by a given entity (like health status or the number of requests issued to a given LBS) and to data that can easily be found in external sources (e.g., the home address) and that, hence, can possibly be used to re-identify a subject. In contrast, the term "private information" specified in the Computer Science literature only refers to information that, intuitively, users are not willing to disclose. So, we can identify two differences:

1. The concept of "non-private attributes", formulated in Computer Science, does not have a counterpart in the legal notion.
2. Private information, as defined in Computer Science, does not include quasi-identifiers while, according to legal definitions, quasi-identifiers are actually considered as personal information.[51]

The consequence of problem (1) is that it can contribute to rendering the solutions proposed in Computer Science not adhering to the legal norm, with a consequent impact, as we shall see in the following, on the applicability and usefulness of the Computer Science solutions.

Probably, one of the reasons that lead to difference (2) is that, from the Computer Science point of view, when the anonymization problem is addressed, it is not necessary to avoid the disclosure of quasi-identifiers since, by definition, the adversary can externally find this information in association with the user's explicit identifier. In practice, it is assumed that if a datum is publicly available, then its re-publication does not violate the subject's privacy. However, this approach does not take into account that from the legal point of view (e.g., in the Italian legal system), even if a datum is already public, it cannot be freely processed, but only be used for the purpose for which it was made public. For example, if personal data on Alice are published in the voters' list, this information cannot be published by a web service for marketing purposes even if there is no additional data associated with Alice's record, unless Alice gives her explicit authorization. In other words, it could be misleading to qualify a datum as "public" because a published datum is not always free from legal constraints. One of the reasons behind this difference is that the concept of "purpose of data processing", which has an important role from the legal point of view, is neglected in Computer Science.

It is worthwhile to wonder whether it is possible to fix the two problems above. For what concerns problem (1), there is an easy way out that consists in assuming that, in each application of the privacy models, the set of "non-private attributes" is empty. The solution to problem (2) is more complicated. Consider the following example.

[51] It is worthwhile to note that some papers that have recently appeared in the computer science literature do not distinguish between quasi-identifiers and personal information. Among others, the paper: Arvind Narayanan and Vitaly Shmatikov, "Robust De-anonymization of Large Sparse Datasets," *IEEE Symposium on Security and Privacy*, 0 (2008): 111-125.

Example 12 In this example we refer to the data reported in Table 4.5. According to the definitions provided in the Computer Science literature, assume that the attributes "Gender", "Date of Birth" and "ZIP Code" are quasi identifiers, while "Disease" is private information. According to the definition of *l*-diversity, the table satisfies the 3-diversity property. However, observe that, if an adversary knows that Alice is one of the subjects in this table, he can discover her year of birth and three digits of her ZIP code (since all women in the dataset have the same values for these attributes). The question is: should the publication of this table (without the explicit user's authorization) be considered a privacy violation? The aim of the models provided in Computer Science is to give an ultimate answer to this question: once a model is defined, it is possible to automatically evaluate whether anonymized information can be published or not. On the other hand, from the legal point of view, a unique answer cannot be provided. It is necessary to take into account the purpose of the publication, the compliance with legal constraints (and the legal constraints differ from one country to another) and the nature of the controller (public or private).

Example 12 shows that, although Table 4.5 satisfies the privacy requirements defined by a privacy model, the disclosure of the table may still be considered non-compliant with the regulations. In other words, the model fails to define when the disclosure does not violate the data respondents' privacy. The technical reason does not lie in a particular problem of the *l*-diversity model, but in a transversal problem that affects most of the privacy models proposed in the literature and in particular their relation to the legal norms. Indeed, some pieces of information can actually be disclosed to those adversaries that have less information than assumed. Consider once again Example 6: an adversary that knows, for each of the subjects in the table, the values of the subject's quasi-identifiers, cannot learn any information from the disclosure of Table 4.5. Vice versa, if an adversary does not know Alice's age, but only that Alice is in that table, he can discover her age. This has an impact on the possibility to disclose Table 4.5. Indeed, despite the fact that Alice's age could be discovered from other sources, according to existing regulations, this datum cannot be freely disclosed.

Technically, a defence technique to contrast the above problem requires to apply an idea similar to the one proposed by Fung[52] (see Sect. 4.3.1.3) that makes it possible to model different quasi identifiers. In practice, instead of considering a single set of quasi-identifying attributes, like in the *l*-diversity model, it would be necessary to model as QI each possible combination of "quasi-identifying attributes". Clearly, it should be investigated whether this approach is practical or not in terms of generalized data quality.

4.4.3 Anonymity Measurement

Another difference between the Legal and Computer Science fields concerns how to evaluate whether an individual is identifiable or not. Note that this topic is of paramount

[52] Id. at 20 ("Anonymizing Classification Data for Privacy Preservation").

importance, since it is needed to evaluate whether data are actually anonymous or can be re-associated with a specific individual.

To the best of our knowledge, one of the main legal references to this problem suggests to measure the difficulty in re-identifying the data subject in terms of "time and manpower".[53] This definition is suitable for traditional computer security problems. For example, the difficulty to decrypt a message without the proper key can be measured in terms of how long would it take to try all possible keys i.e., the so called "brute force" attack. However, the question is: does the same measure apply to the problem of guaranteeing privacy? As shown in Sect. 4.3, all the formal models proposed in the Computer Science literature indicate that the key factor affecting the difficulty to re-identify an anonymous datum is the background knowledge available to the adversary, while the adversary's manpower and time to perform the attack are not relevant parameters. Consider, for instance, Table 4.4 in Example 5. Even if the adversary has almost infinite resources (computational power, time and manpower), it would not be possible to identify Shunsuke's data record to infer his disease without additional information. Vice versa, if the adversary knows a piece of background knowledge as in Example 5, i.e., Shunsuke is in the database, was born in 1982 and is Japanese, then it is easy to immediately infer that Shunsuke has a cold, even with negligible computational power, time and economic resources.

According to the above consideration, it seems more reasonable that "time and manpower" should not be adopted to directly measure the effort required to violate anonymity but that, instead, they should measure the effort required by the adversary to acquire background information that in turn can be used to re-identify a data subject. For example, the knowledge of the adult individuals living in a certain area, together with some personal information (e.g., date of birth, home address, etc.…) should be considered as "reasonably" available information for any adversary, since this information is contained in the voters list that in many countries can be obtained for free or at a small price. Vice versa, the "full background location knowledge" (see Sect. 4.3.2.2) could be obtained by physically spying a set of persons or, with some additional approximation, by violating the information system of mobile phone operators, hence acquiring the traces of movements of a large number of users. Both solutions for acquiring the "full background location knowledge" would probably be considered as "unreasonably costly".

It would therefore be desirable, under the legal point of view, to clarify the notion of reasonableness, taken as a measurement criterion of time, cost and resources. We believe that this clarification should be one of the main purposes of the next reform of the European Directive on personal data protection. In this respect, we suggest that reasonableness should be intended as "reasonableness of knowledge" by third parties of information and criteria for the identification of subjects.

[53] Id. at 13 ("Recommendation No. R (97) 5 on the protection of medical data").

4.4.4 Anonymity and the Principle of Minimization

According to the principle of minimization, personal data processing is allowed only for the achievement of a specified purpose and, if this task can be accomplished with anonymous or pseudonymous data, this form of information should be preferred. The objective of this principle is to promote the use of anonymous or pseudonymous data when possible. However, as we shall see in the following, some technical problems arise in the application of this principle.

In many cases transforming data to achieve anonymity causes information loss, and this can make the result of the subsequent analysis approximate. Consider for instance a research centre that wants to know the date of birth of the users for each ZIP code value. If this query is performed using the exact data (e.g., Table 4.1), the answer contains the exact dates of birth. In contrast, if the query is performed on the data in Table 4.2, the research centre can only know the year of birth. Clearly, the result of the query in this last case is less accurate, but in some contexts it could be acceptable if, at the same time, it does not reveal the data subject's personal information.

The problem here is the following: the process of rendering the information anonymous, as commonly intended in the Computer Science literature, necessarily involves a form of data suppression and/or generalization. This implies that the resulting information is less accurate than the original one. Consequently, in many cases, the anonymous version of the information makes it impossible to achieve exactly the same results that would be achieved with non-anonymous data, hence motivating the disclosure of the non-anonymous information. In other words, since the principle of minimization does not take into account any form of approximation in the result, it can be used as a motivation for a controller not to release anonymous data, which is conceptually opposes the core idea behind the principle of minimization.

One final observation: the minimization principle is general and, in itself, must be shaped case by case. Indeed there may be situations in which the value of information plays a predominant role with respect to its "confidentiality". However, this does not apply in general. Perhaps a specification of this principle, or simply a reinterpretation of this principle, in light of the standard of reasonableness, would enhance its practical applicability.

4.5 Conclusions and Future Work

In this paper we addressed the topic of anonymity as a tool to protect personal privacy. The overall objective was to encourage the discussion between Law and Computer Science experts on a topic that is bound to be subject of research in the next years. To achieve this, we presented a brief analysis of the state of the art of this problem from the two points of view. Despite the different methodological

approaches, the challenge was to identify a common language for general definitions. This highlighted the fact that some notions commonly adopted in the Computer Science literature do not find any legal support. Analogously, some legal definitions seem to ignore conceptual issues that are clearly identified in the formal models proposed in Computer Science research. Overall, this paper identifies a few common aspects and several differences between the definitions and results suggested in the two disciplines.

In particular, we observed that the notion of anonymity has a central role both in regulations on personal data protection and in the techniques proposed to protect subject's privacy. Indeed, the anonymity measures proposed in the Computer Science field, support the fact that anonymity is a relative notion that depends on the context. On the other hand, Computer Science has shown the limits of anonymity, hence posing new juridical questions about its role. Despite this point in common, an agreement is missing on some of the basic concepts related to anonymity, like the notion of quasi-identifiers and personal data. This poses new challenges to researchers in both communities. Similarly, according to the state of the art in the two areas, it is still unclear how to measure the "level of anonymity" of a datum. If the interpretation of European legislation suggested in Sect. 4.4 is accepted, and the problem is clarified under the legal point of view, it will be necessary to identify the most suitable formal models to practically compute the measure. Finally, we considered the principle of minimization, showing how its current formulation can motivate the processing and disclosure of identified information, in contrast with the overall idea of this principle.

This paper poses the basis for a new approach to the analysis of the personal data protection problem, suggesting a number of new challenges and research directions.

First of all we plan to extend research to the general problem of privacy protection beyond anonymity. Indeed, there are some concepts that need to be investigated, including the legal foundations of the "obfuscation" functions (see Sect. 4.4.1) and the involved privacy "negotiation" between the controller and subject. Another topic, which is becoming popular in the Computer Science community, is the notion of "differential privacy": it would be of great interest to analyze this concept from the legal point of view, making an effort to identify whether it is compliant with the law. Moreover, it could be interesting to analyse the anonymity problem in "credential systems" in which each user is identified by a different pseudonym by different organizations. The challenge is to prevent the possibility to link different pseudonyms.[54]

Another research effort should be devoted to analyzing the existing privacy protection tools available in commercial applications and services. Indeed, in absence of consolidated technical solutions based on sound legal bases, business companies are addressing the personal data protection problem with ad hoc solutions, and in some case it can be unclear which are the technical or legal fundamentals of these techniques.[55]

[54] David Chaum, "Showing credentials without identification transferring signatures between unconditionally unlinkable pseudonyms," in *Advances in Cryptology - AUSCRYPT '90*, 453:245-264, Springer Berlin/Heidelberg, 1990.

[55] This problem can also be focused in the discussion about on the notion of "accountability".

Considering the privacy problem from a practical point of view, the topic of privacy preservation in social networks would definitively deserve a thorough investigation with the interdisciplinary methodology adopted in this paper. Indeed, although it has already been recognized that specialized techniques are required for these specific services, it is still unclear whether the existing norms can be adapted to this context. For example, one problem is that each data subject can publish information about other users, hence playing the role of the controller. In general, these services involve at the same time categories of subjects having different roles with respect to the processing of data, and it is unclear whether these subjects are captured by existing legal norms. Vice versa, it is necessary to have a clear mapping of the roles involved in the data processing and of the connected liabilities.

As we observed, there are several open issues that need to be addressed. Consequently, it is necessary to continue and enhance the dialogue between researchers in the Law and Computer Science communities, in order to allow the possibility of satisfying the need to balance the use of advanced technologies with the protection of individual fundamental rights. The necessity to develop shared solutions to this problem is part of a process that cannot be anything but interdisciplinary. Indeed, without a practical approach, the risk is that Law becomes hardly applicable. Analogously, Computer Science risks to be a dead end if it is not modelled according to the regulations in force.

References

Agrawal, Rakesh, and Ramakrishnan Srikant. 2000. Privacy-preserving data mining. In *Proceedings of the 2000 ACM SIGMOD international conference on management of data*, 439–450. New York: ACM.

Bettini, Claudio. 2010. Privacy and anonymity in location data management. In *Privacy-aware knowledge discovery: Novel applications and new techniques*, ed. F. Bonchi and E. Ferrari. Boca Raton: Chapman & Hall/CRC Data Mining and Knowledge Discovery Series.

Bettini, Claudio, Sergio Mascetti, X. Sean Wang, and Sushil Jajodia. 2007. Anonymity in location-based services: Towards a general framework. In *Proceedings of the 2007 international conference on mobile data management*, 69–76. Washington, DC: IEEE Computer Society.

Chaum, David. 1990. Showing credentials without identification transferring signatures between unconditionally unlinkable pseudonyms. In *Advances in Cryptology – AUSCRYPT '90*, ed. J. Seberry, J. Pieprzyk, 453:245–264. Berlin/Heidelberg: Springer.

Ciriani, Valentina, Sabrina di Vimercati, Sara Foresti, and Pierangela Samarati. 2007. Microdata protection. In *Secure data management in decentralized systems*, vol. 33, ed. Yu Ting and Sushil Jajodia, 291–321. New York: Springer.

Directive 95/46/EC of the European Parliament and of the Council of 24 October 1995 on the protection of individuals with regard to the processing of personal data and on the free movement of such data, OJ L 281, 23.11.1995, 31–50.

Directive 2002/58/EC of the European Parliament and of the Council of 12 July 2002 concerning the processing of personal data and the protection of privacy in the electronic communications sector (Directive on privacy and electronic communications), OJ L 201, 31.7.2002, 37–47.

Dwork, Cynthia. 2006. Differential privacy. In *Automata, languages and programming*, 4052:1–12, ed. Michele Bugliesi, Bart Preneel, Vladimiro Sassone, and Ingo Wegener. Berlin/Heidelberg: Springer.

Finocchiaro, Giusella. 2009. Anonymity and the law in Italy. In *Movement-aware applications for sustainable mobility: Technologies and approaches*, ed. Ian Kerr, Valerie M. Steeves, and Carole Lucock, 523–536. Oxford: Oxford University Press.

Finocchiaro, Giusella, and Claire Vishik. 2010. Law and technology: Anonymity and right to anonymity in a connected world. In *Movement-aware applications for sustainable mobility: Technologies and approaches*, ed. Monica Wachowicz, 140–156. Hershey: IGI Global.

Fung, Benjamin C.M., Ke Wang, and Philip S. Yu. May 2007. Anonymizing classification data for privacy preservation. *IEEE Transactions on Knowledge and Data Engineering* 19(5): 711–725.

Gruteser, Marco, and Dirk Grunwald. 2003. Anonymous usage of location-based services through spatial and temporal cloaking. In *Proceedings of the 1st international conference on mobile systems, applications and services*, 31–42. MobiSys '03. New York: ACM.

Italian Personal Protection Code, Legislative Decree no. 196, 30/06/2003, art. 4, co. 1, lett. n.

Kalnis, Panos, Gabriel Ghinita, Kyriakos Mouratidis, and Dimitris Papadias. December 2007. Preventing location-based identity inference in anonymous spatial queries. *IEEE Transactions on Knowledge and Data Engineering* 19(12): 1719–1733.

Li, Tiancheng, and Ninghui Li. 2009. On the tradeoff between privacy and utility in data publishing. In *Proceedings of the 15th ACM SIGKDD international conference on knowledge discovery and data mining*, 517–526. New York: ACM.

Li, Ninghui, Tiancheng Li, and S. Venkatasubramanian. 2007. *t*-closeness: Privacy beyond *k*-anonymity and *l*-diversity. In *IEEE 23rd international conference on data engineering, 2007* (*ICDE 2007*), 106–115. Istanbul, Turkey: IEEE Computer Society.

Machanavajjhala, Ashwin, Daniel Kifer, Johannes Gehrke, and Muthuramakrishnan Venkitasubramaniam. March 2007. *l*-diversity: Privacy beyond *k*-anonymity. *ACM Transactions on Knowledge Discovery from Data* 1(1): 24.

Mascetti, Sergio, Claudio Bettini, X. Sean Wang, and Sushil Jajodia. 2006. *k*-anonymity in databases with timestamped data. In *Proceedings of the thirteenth international symposium on temporal representation and reasoning*, 177–186. Washington, DC: IEEE Computer Society.

Mascetti, Sergio, Claudio Bettini, Dario Freni, and X. Sean Wang. September 2007. Spatial generalisation algorithms for LBS privacy preservation. *Journal of Location Based Services* 1(3): 179–207.

Mokbel, Mohamed F., Chi-Yin Chow, and Walid G. Aref. 2006. The new casper: Query processing for location services without compromising privacy. In *Proceedings of the 32nd international conference on very large data bases*, 763–774. VLDB '06. Seoul, Korea: VLDB Endowment.

Monreale, Anna, Dino Pedreschi, and Ruggero G. Pensa. 2010. Anonymity technologies for privacy-preserving data publishing and mining. In *Privacy-aware knowledge discovery: Novel applications and new techniques*, ed. F. Bonchi and E. Ferrari. Boca Raton: Chapman & Hall/ CRC Data Mining and Knowledge Discovery Series.

Narayanan, Arvind, and Vitaly Shmatikov. 2008. Robust de-anonymization of large sparse datasets. In *Proceedings of 29th IEEE symposium on security and privacy*, vol. 0, 111–125. Los Alamitos: IEEE Computer Society.

Ohm, Paul. 2009. Broken promises of privacy: Responding to the surprising failure of anonymization. *UCLA Law Review* 57:1701, 2010.

Opinion 4/2007 of the Article 29 data protection working party on the concept of personal data, WP 136, 20.06.2007.

Recommendation No. R (97) 5 of the Committee of Ministers to Member States on the protection of medical data, 13/02/1997.

Recommendation No. R (97) 18 of the Committee of Ministers to Member States on the protection of personal data collected and processed for statistical purposes, 30/09/1997.

Riboni, Daniele, Linda Pareschi, Claudio Bettini, and Sushil Jajodia. 2009. Preserving anonymity of recurrent location-based queries. In *Proceedings of the 16th international symposium on temporal representation and reasoning*, 62–69. TIME '09. Washington, DC: IEEE Computer Society.

Samarati, Pierangela, and Latanya Sweeney. 1998. Generalizing data to provide anonymity when disclosing information (abstract). In *Proceedings of the seventeenth ACM SIGACT-SIGMOD-SIGART symposium on principles of database systems*, PODS '98. New York: ACM.

Schwartz, Paul M., and Daniel J. Solove. 2011. The PII problem: Privacy and a new concept of personally identifiable information. *New York University Law Review* 86: 1814–1894.

Terrovitis, Manolis, and Nikos Mamoulis. 2008. Privacy preservation in the publication of trajectories. In *Proceedings of the ninth international conference on mobile data management*, 65–72. Washington, DC: IEEE Computer Society.

Winkler, William E. 1999. *The state of record linkage and current research problems*. Washington, DC: Statistical Research Division, U.S. Bureau of the Census.

Working Party document on data protection issues related to RFID technology, WP 105, 19/01/2005, Art. 8.

Xiao, Xiaokui, and Yufei Tao. 2006. Personalized privacy preservation. In *Proceedings of the 2006 ACM SIGMOD international conference on management of data*, 229–240. SIGMOD '06. New York: ACM.

Yakowitz, Jane. 2011. Tragedy of the data commons. *Harvard Journal of Law and Technology 25(1)*, Fall 2011.