Some basics on privacy techniques, anonymization and their big data challenges

Julián Salas and Josep Domingo-Ferrer

Abstract. With the progress in the information and communication fields, new opportunities and technologies for statistical analysis, knowledge discovery, data mining, and many other research areas have emerged, together with new challenges for privacy and data protection.

Nowadays several personal records are kept in computerized databases. Personal data is collected and kept in census databases, medical databases, employee databases, among others. There has always been an asymmetry between the benefits of computerized databases and the rights of individual data subjects. Some data protection principles can be derived from the legal framework.

In this survey, we present some basic cryptographic and non-cryptographic techiques that may be used for enhancing privacy, we focus mainly on anonymization in databases and networks, discuss some differences and interactions among the well known models of k-anonymity and differential privacy and finally present some challenges to privacy that come from big data analytics.

1. Introduction

Privacy is a fundamental human right, United Nations Declaration of Human Rights (UDHR) 1948, Article 12 states: "No one shall be subjected to arbitrary interference with his privacy, family, home or correspondence, nor to attacks upon his honour and reputation. Everyone has the right to the protection of the law against such interference or attacks."

However, there has always been an asymmetry between the benefits of computerized databases and the rights of individual data subjects. In an effort to balance such an asymmetry, the Code of Fair Information practices was published as a central part of the report of the Committee of the Secretary of Health, Education, and Welfare, Records, Computers, and the Rights to Citizens [74]. It enunciates five fundamental principles in record keeping which are: the prohibition of secret databases, data subjects must be allowed to inspect their records and how are they used, the data obtained for one purpose may not be used for other purposes without the consent of the data subject, the data subject must be able to correct or amend her/his records, and the data must be kept reliable and secure.

Data protection principles can be derived from the legal framework. According to European data protection law, the processing of personal data is legitimate if: the individual whose personal data are being processed (the data subject) has unambiguously given consent, or processing is necessary for the performance of a contract, for compliance with a legal obligation, for protecting vital interests of the data subject, for the performance of a task carried out in the public interest, or for the purposes of legitimate interests pursued by the data processing entities except when such interests are overridden by the fundamental rights and freedoms of the data subject.

In [39], based on the legal framework the following eight privacy by design strategies are defined and explained: *Minimise, Hide, Separate, Aggregate, Inform, Control, Enforce and Demonstrate.* The amount of personal information processed should be minimal, it should be hidden from plain view, the processing should be done in distributed fashion whenever possible, personal information should be processed at the highest level of aggregation with the least possible detail in which it is still useful, data subjects should be informed whenever personal information is processed, and they should have agency over the processing of their personal information, a privacy policy compatible with legal requirements should be enforced, and the data controller should be able to demonstrate compliance with the privacy policy and legal requirements.

As [53] points out: with the advances in digital media, we have witnessed a dramatic rise in monitoring and tracking by technical means, which has lead to a shift in its nature. It has been automated, undiscriminating and has incorporated new subjects, monitors and motives. These transformations have affected the state and practice of electronic engagement with personal information, which, in turn, are experienced as threats to privacy.

Following these and other concerns and guidelines, several technical means have been developed for providing privacy.

In this survey, we present some cryptographic and non-cryptographic techiques that may be used for enhancing privacy, we focus mainly on anonymization in databases and networks, discuss some differences and interactions among the well known models of k-anonymity and differential privacy and finally present some challenges to privacy that come from big data analytics.

2. Cryptographic and non-cryptographic techniques for enhancing privacy

In this section we present a brief list of cryptographic and non-cryptographic techniques for enhancing privacy with a short explanation for each of them. For a more thorough explanation, cf. [21].

2.1. Cryptographic techniques

- *Authentication* is usually the first step in using a remote service, consists of methods to confirm that a user is actually who claims to be, then secure communications can proceed based on parties knowing each other's identity.
- *Attribute based credentials* allow a user, given a set of attributes, to securely and privately prove ownership of these attributes to a verifier, disclosing some of them while keeping the others secret. A credential is signed by a trusted issuer that guarantees to provide valid values for the attributes in the credential. The attributes contained in a credential cannot be forged, the use of a credential cannot be linked to a former case where it was used, the credentials cannot be transferred to different subjects, can be revoked and in case of abuse the identity of the user can be recovered.

  Two ways for obtaining unlinkable credentials are to use credentials only once, and use a Blind Signature Protocol, or use them several times and use Zero Knowledge Proofs for obtaining unlinkability.
- *Secure private communications* must be established for different services such as VoIP, e-mail or social networking. This may be done by encrypting the communications between users in an end-to-end fashion, making the content of communications unintelligible to any third parties.
- *Communications anonymity and pseudonymity* protect the meta-data of the communication, that may be: who communicates with whom, the time and volume of messages, the location and the identity of the network end-points.

  The basic means to achieve communications anonymity is the use of proxies or VPNs. Depending on a single operator may facilitate an attacker to observe or coerce it to reveal the identities of communicating parties, hence Onion routing [35] uses multiple relays. However, it is vulnerable to a global passive adversary that may use statistical analysis. Mix networks

[12], communicate with messages of a uniform measure, allow for delays and thus prevent such attacks.

All systems providing anonymity properties make use of messages from multiple users in a way that is difficult for the adversary to distinguish them. As a result, anonymous communications systems benefit from large volumes of users, as many other privacy notions.

- *Storage privacy* is achieved commonly by user authentication and access control lists. Also by storing the data in encrypted form, which can be done by full disk encryption (FDE), file system level encryption (FSE), or it can be done in a steganographic file system [2], which hides also the existence of the encrypted data. When encrypted and steganographic storage is used in remote settings like in cloud storage, for searching such data, the techniques of Symmetric Searchable Encryption [18] and Public-key Searchable Encryption [6] are used.
- *Privacy preserving computations* can be achieved by means of homomorphic encryption, which allows for computations to be carried on cyphertexts. Privacy homomorphisms were introduced in [55] and were broken by ciphertext-only attacks or known-cleartext attacks [9]. Partially homomorphic cryptosystems (that allow for only one operation) include RSA [56], ElGamal [31] and Paillier cryptosystem [51] among others. Fully Homomorphic Encryption (FHE) [33], on the other hand, supports arbitrary computation on ciphertexts and is far more powerful.

  Secure multi-party computation methods, which enable several parties to jointly compute a function over their inputs, while keeping these inputs private, were introduced in [69]. Two primitives for building multi party computation protocols are oblivious transfer and secret sharing. An oblivious transfer protocol [54] consists of a sender which transfers one of potentially many pieces of information to a receiver, but remains oblivious as to what piece has been transferred. Secret sharing [4, 66] refers to methods for distributing a secret amongst a group of participants. The secret can be reconstructed only when at least a given number of shares are put together.
- *Privacy in databases* may refer to privacy of the respondents to a statistical database, to the databases owned by different corporations, or to the users that access to a resource. Hence it may be divided in the following three categories: owner privacy, user privacy and respondent privacy.

  Owner privacy has been attained by privacy preserving data mining, which considers the results of the data analysis and knowledge extracted, such as the rules in association rule mining which can be considered sensitive [68] or building a decision tree classifier in which the training data values of individual records have been perturbed [1]. Other approaches consider computing a data mining algorithm on the union of the databases of several parties without revealing more information to others than the output of the computation, such as [47], in which a decision tree is jointly calculated by two parties, using Secure Multi-party Computation.

  User privacy has been obtained mainly by cryptographical protocols as in Private Information Retrieval [14].

2.2. Non-cryptographic techniques

- *Transparency enhancing techniques* have been proposed to empower the users to understand what data and for which purpose is collected and processed and the relation of such data to their privacy. A survey can be found in [38].
- *Intervenability* provides control by the users over the data processing, it allows their intervention when needed, it is also related with the rights to rectification and erasure of data and the right to withdraw consent.
- *Privacy in databases:*

  Other approaches for *user's privacy* (such as [22, 40]) avoid profiling by the search engine by adding noise to the queries sent. Or, dissociate the different facets of an individual, by creating one virtual identity for each user's facet ([41]), obtaining good query results by

allowing profiles for each facet and privacy by separating each of the interests that together could be used for profiling.

Respondent privacy has been pursued by statisticians and computer scientists working in statistical disclosure control (SDC), also known as statistical disclosure limitation (SDL). It is the topic of the following section.

3. Statistical Disclosure Control methods for microdata protection

3.1. Microdata protection

Microdata contain a set of attributes relating to single respondents in a sample or in a population, can be represented as tables composed of tuples (records) with values from a set of attributes, which may be numerical (e.g. age, blood pressure) or categorical (e.g. gender, level of studies, job). Attributes can be classified as Identifiers, Quasi-Identifiers, Confidential, and Non-confidential.

It was traditionally believed that anonymization could be achieved by only removing all the *Identifiers*, that is, attributes that unambiguously identify respondents (such as Social Security Number, Passport, Name-surname). However, unique combinations of attribute values may be used to reidentify unambiguously an individual in a database without explicit identifiers. Hence those attributes are called quasi-identifiers (QIs).

Since QIs can be used to relate anonymized records to external non-anonymous databases, this may lead to re-identification. And by reidentifying a respondent of a database her *Confidential* (sensitive) attributes may be revealed (e.g., Salary, Medical conditions, etc.) Therefore, anonymization techniques must deal with QIs. These techniques are based on reducing the amount or precision of the released information, and is achieved by two main principles, masking the data (releasing a modified version of the original dataset) or by releasing synthetic values instead of the real ones.

Masking can be divided in two categories: non-perturbative and perturbative. Non-perturbative masking reduces the level of detail of the original data but does not distort it.

Some well known *non-perturbative masking techniques* are: publishing only a sample of the original data file (*Sampling*); coarsening a categorical attribute by combining several categories to a more general one (*Generalization*) or replacing numerical values by intervals; putting values above/below a given threshold into a single category (*Top/Bottom coding*) or suppresing certain values of individual attributes in order to decrease the uniqueness of the individual (*Local suppression*).

Some *perturbative masking techniques* such as *noise addition, rank swapping, post-randomization and microaggregation* are special cases of matrix masking [30], where the original microdata set is X and the masked microdata set is Z computed as $Z = AXB + C$, matrix A is called a record-transforming mask, matrix B an attribute transforming mask and matrix C a displacing mask.

For an overview over the different algorithms for *noise addition*, see [8].

In *Data swapping* the attributes values of some individual records are exchanged in such a way that marginals are maintained. *Rank swapping* is a variant of data swapping that can be applied to numerical attributes. In the *Post-randomization* method the values of categorical attributes are changed according to a Markov matrix. *Microaggregation* consists on partitioning the records into groups containing each at least k records and publishing the average record of each group. In [25] microaggregation was applied for obtaining k-anonymity.

3.2. Two main models for privacy protection

The two main models for privacy protection, from which many others have been developed, are k-anonymity and ϵ-differential privacy.

Both have a priori guarantees on the disclosure risk by definition. On the one hand, k-anonymity implies that an individual cannot be reidentified with probability greater than $\frac{1}{k}$. On the other hand, ϵ-differential privacy implies that not even the presence of an individual may be guessed by analyzing the anonymized dataset.

The concept of *k-anonymity* was defined to release personal data while safeguarding the identities of the individuals to whom the data refer [59]. A dataset is k-anonymous if each record is indistinguishable from at least other $k-1$ records within the dataset, when considering the values of its QIs. This guarantees that the individuals cannot be re-identified by linking attacks with probability less than $\frac{1}{k}$. However, when the sensitive attributes on a group of k-anonymous records, have low variablity (e.g., when they are all equal), there is no need of reidentification to disclose the value of the sensitive attribute of a record. This remark was done in [52], who proposed the model of ℓ-diversity for solving this issue.

A k-anonymous data set is said to be *ℓ-diverse* if, for each group of records sharing quasi-identifier values, there are at least ℓ well-represented values for the sensitive attribute. Later, in [44] it was shown that the model of ℓ-diversity does not prevent attribute disclosure when the overall distribution of the sensitive attribute is skewed. Hence, they proposed the t-closeness model.

A k-anonymous data set is said to have *t-closeness* if, for each group of records sharing quasi-identifier values, the distance between the distribution of each sensitive attribute within the group and the distribution of the attribute in the whole data set is no more than a threshold t. In general, k-anonymity may be obtained by perturbative or non-perturbative means, it can be obtained by generalization and suppression or by microaggregation.

A randomized function $\mathcal{K}$ gives *ϵ-differential privacy* if for all data sets D_1 and D_2 differing on at most one element, and all $S \subseteq Range(\mathcal{K})$,

$$Pr[\mathcal{K}(D_1) \in S] \leq exp(\epsilon) \times Pr[\mathcal{K}(D_2) \in S]. \tag{1}$$

Differential privacy establishes that the removal or addition of a single element in the database does not (considerably) change the results of an analysis. Therefore, the presence or absence of any individual on the dataset is not revealed (up to $exp(\epsilon)$) by the computation.

For any given function f, its *global sensitivity* $S(f)$ measures the maximum difference in the answer to that query over all pairs of neighboring data sets [27]. If we denote by Lap(λ) the Laplace distribution, then, the mechanism $\mathcal{K}_f(x) = f(x) + Lap(S_f(\epsilon))$ is ϵ-differentially private.

To provide differential privacy for non-numerical queries, the *exponential mechanism* may be used [50].

A relaxation of ϵ-differential privacy is to allow the outputs that violate inequality (1) to occur with a small error probability δ, then:

A randomized function $\mathcal{K}$ gives *(ϵ, δ)-differential privacy* if for all data sets D_1 and D_2 differing on at most one element, and all $S \subseteq Range(\mathcal{K})$,

$$Pr[\mathcal{K}(D_1) \in S] \leq exp(\epsilon) \times Pr[\mathcal{K}(D_2) \in S] + \delta. \tag{2}$$

Nonetheless, it has been observed that to provide strong privacy protection, it must hold that $\delta \ll \frac{1}{n}$. An example from [45], is the following:

The δ-sampling algorithm $\mathcal{K}_\delta$ that goes through each tuple t in the input database, publishes it with probability δ, and suppresses it with probability $1-\delta$.

Note that $\mathcal{K}_\delta$ satisfies $(0, \delta)$-differential privacy. Nonetheless, if there are n tuples in the database, $\mathcal{K}_\delta$ publishes on average $n\delta$ tuples unchanged. It can be concluded that (ϵ, δ)-differential privacy guarantees that for any tuple, a privacy breach can occur with probability at most δ. Hence, when there are many tuples, the probability that the privacy of some tuple will be breached increases.

An ideal privacy definition for SDC was stated in [20], it says that anything about an individual that can be learned from the database can also be learned without accessing the database. However,

[28] showed that this is not possible. Certainly this property may be satisfied if the mechanism provides no information, but we must recall that the objective of releasing data is to provide utility. Therefore, this original notion of zero-knowledge (adopted from cryptography) was relaxed and used for modeling privacy in social networks in [32] to allow the mechanism to release some aggregate information that is considered acceptable to release but not any additional information beyond this.

3.3. Some differences and interactions between these two models

If the goal is to publish data without making any assumptions on the analyses or queries that will be carried on the data, then the anonymized data may be published and thus will be available for any type of analysis.

In contrast, in the interactive setting, the analysis should be known in advance in order to calibrate the sensitivity of the query to the parameter ϵ. The number of queries may be subject to a pre-specified privacy budget ϵ; otherwise, as shown in an extreme counter-example, asking the same query repeatedly and taking the average of the responses will eventually reveal the true answer.

Releasing a dataset satisfying k-anonymous (or any kind of syntactic model) has the problem that a previously unknown QI may appear, or even the separation between QI and sensitive attributes may not be such; it may be the case that sensitive attributes may be used as quasi-identifiers.

It is easier to interpret the relation of the value of parameter k with the probability of identification than the relation of parameter ϵ. However, in both cases, it may be interesting to carry out an ex-post analysis of the disclosure risk.

Given the criticisms made to both syntactic anonymity and differential privacy, they may seem to be opposed one to the other. However, as it has been observed by [16] both approaches have their place; each approach has issues that should be addressed by further research. The sytactic model (k-anonymity and its variants) may be better suited for privacy-preserving data publishing while differential privacy for privacy-preserving data mining. Morever, there are positive interactions among these two models: for example, [46] relates k-anonymity with differential privacy, [23] relates t-closeness with differential privacy and [61] provides a method in which k-anonymity is used for improving the utility of differential privacy. Finally, it seems that an ex-post analysis is still necessary when anonymizing a specific dataset in such a way to evaluate the guarantees of each method with specific parameters. A unifying view of the different methods in which empirical utility and privacy are defined is [17].

3.4. Comparing different methods: Utility loss and disclosure risk for SDC methods

There is always a trade-off between the risk of disclosure and the utility loss. To achieve a higher level of privacy a larger amount of modifications must be done to the data, therefore increasing the damage on the utility.

Utility loss can be evaluated using generic or specific measures. Some examples of generic measures are the collection of basic statics such as means, covariances, etc., see [26]. Specific measures analyse the impact of an anonymization on the output of a given analysis.

The disclosure risk in microdata can be assessed by record linkage techniques, cf. [24]. In distance based record linkage, the original and the masked data sets are compared based on a predefined distance. The masked record is assigned to its nearest record in the original data, and is labeled as correctly linked when the original and masked record are the corresponding ones. Then, the percentage of correctly linked records is a measure of disclosure risk.

A tool for comparing different methods are the R-U maps [29], that consist of plotting the measure of data utility on one axis and the disclosure risk on the other. In this way, a clearer exposition of the tradeoffs between different methods and parameters can be displayed, which facilitates their evaluation.

4. Online Social Network Anonymization

Online social networks have become a part of the daily lives of most people. They reflect the interests and relations of their participants, hence they provide a great oportunity to analyze them and extract information that may be valuable for the benefit of our society.

Moreover, networks can represent very diverse objects other than social networks, such as the structure of an organization, distribution networks, neural or metabolic networks. They may be labeled, or weighted, and directed to represent the characteristics of the given nodes or edges, that may be the strength and direction of the relation (edge), the size or any other property of an object (node). Nodes may represent very different objects such as individuals, locations, enzymes and metabolites, and edges may represent diverse relations such as, acquaintances, communication, geographical proximity, interaction among many others.

In the case of online social networks such as in statistical databases, publishing them for their study yields the possibility of knowing personal characteristics of their members, with the risk of using them to reidentify the individuals behind the nodes and revealing their private attributes.

This implies the need for modification of characteristics that may lead to reidentification or to revealing attributes that may be considered private.

4.1. Graph anonymization

Although the concepts of network and graph have usually the same meaning, we will use graph to emphasize that it is only the structural part and network when considering possibly additional characteristics of the individuals, that may be demographics, likes or any other interesting properties of the network.

Backstrom et al. [3] showed that simple anonymization (only removing or replacing names or identifiers) of a social network may still allow an attacker to learn relations between targeted pairs of nodes and reidentify the original nodes.

They showed that the structural properties (the graph) of a social network can be used for reidentification. Hence, the methods and concepts of privacy from statistical disclosure control have been considered for graphs, such as k-anonymity [60, 65].

In the survey [72] graph anonymization strategies are characterized as clustering based or graph modification approaches. The modification based approaches anonymize a graph by inserting or deleting edges or vertices, which can be done in a greedy or a randomized way.

Clustering approaches for graph anonymization consider clustering edges as in [70] or vertices such as [37, 10]. Hay et al. [37] developed a randomization method by changing the original graph with m edge deletions and m insertions chosen uniformly at random, and calculating the protection it provides. This can be related to noise addition in SDC.

4.1.1. k-anonymity for graphs. The concept of k-anonymity has several different definitions for graphs depending on the assumption on the attacker's knowledge, e.g., k-degree anonymity [48], k-neighborhood anonymity [71]. In general all of them can be viewed as k-Candidate anonymity [37] or $k - \mathcal{P}$-anonymity [13], i.e., for a given property $\mathcal{P}$ and a vertex in the graph G there are at least $k - 1$ other vertices with the same property $\mathcal{P}$.

Some examples of structural properties relevant for social network analysis are the *centrality measures*, such as betweenness: the fraction of shortest paths between any two nodes u and v that pass through the vertex in question over all possible shortest uv-paths (i.e., to what extent an individual lies between other individuals); closeness: the average length of the shortest path between the node and all other nodes in the graph (i.e., to what extent the node is near all other individuals); degree: the number of relationships to other individuals in the network. Other examples of measures are the distances between pairs of vertices in the network or whether one is reachable from the other by following a path, or the clustering coefficient of a node that measures the proportion of vertices in its neighborhood that are connected.

All of the above metrics may be considered private for some individuals.

In this case $\mathcal{P}$ is considered the attacker's knowledge, and note that when a graph is $k-\mathcal{P}$-anonymous and $\mathcal{P}$ is the neighborhood, as in [64], any other possible structural property $\mathcal{P}'$ will be $k-\mathcal{P}'$-anonymous.

On the other hand, most of these definitions imply k-degree anonymity. Therefore the minimum number of edge modifications needed to obtain k-degree anonymity may be used as a lower bound for all the other properties.

k-degree anonymity has the additional restriction that the k-anonymous degree sequence must be graphic, i.e., it must correspond to the sequence of a graph. Additional conditions for degree sequences to be graphic and for applications to k-degree anonymization can be found in [57, 58]. These conditions (namely P-Stability) may also be used for edge randomization while preserving the degree sequences.

4.1.2. Differential privacy for graphs. The property of differential privacy, that the outcome of a query is not affected if an individual opts-out or is included in the database, is equivalent to node-differential privacy in the context of online social networks, considering that individuals are usually represented by nodes.

Hence, node-differential privacy considers two graphs G_1 and G_2 that correspond to D_1 and D_2 from Equation (1) such that one can be constructed from the other by removing a node and all of its incident edges.

However, it has been observed that node-differentially private algorithms based on local sensitivity return query answers that are too noisy for practical applications, e.g., [36, 42].

Two neighboring networks (that differ in only one node) may have quite different properties. For example, one may have a huge diameter, say 10^6 and by adding a node connected to all the others it will change its diameter to 2, or the other way around.

Therefore, the concept of edge-differential privacy has been suggested, in which G_1 and G_2 are considered to differ only in one edge.

Other relaxations of node-differential privacy have been studied, such as k-edge differential privacy in [36] which considers that two graphs G_1 and G_2 are neighbors if they differ in at most k-edges. Or considering the projection to graphs with bounded degree as in [42] and [5].

However, when considering these restrictions on the definitions of differential privacy we may no longer keep the original promise of differential privacy to make no assumptions about the kind of attacks against the released statistics nor about the additional information the attacker might possess. A key challenge in the differentially private analysis of social networks is that, for many natural queries, both global and smooth sensitivity can be very large, as it can be seen in the examples in [5].

5. Big data privacy challenges

The aims of privacy protection and personalization in big data analytics may seem to be opossite. Privacy protection, in particular anonymization may be seen as an obstacle to big data analytics.

Big data represents a great opportunity for improving our knowledge as a society and as individuals. It is often described by three Vs of volume, variety and velocity, and has been extended to five Vs adding the variability and value. According to IBM 2.5 quintillions of bytes of data are created every day. This summed up to 4 zettabytes of data worldwide in 2013 [49]. Real time streams of data are collected by sensors that gather climate information, posts on social media sites, purchase transaction records, and cell phone signals among many others. The World Economic Forum [73] suggests that we are at the beginning of the data revolution era, and that this will impact all aspects of our society.

As Giannotti et al. state in [34], one of the big data uses may be that of social sensing, for which the aim is to understand human behavior by leveraging individual profiles, analyzing collective behaviors and social relations, for studying social contagion, spreading of epidemics or opinion and sentiment evolution. However, to protect the data subjects rights and promote the participation and dissemination, trusted networks and privacy-aware mining must be pursued and anonymization methods for such data must be developed.

A thorough explanation of the moral reasons for privacy protection and the different threats to which it is exposed in the information society can be found in [67]. It is also suggested that privacy protection should be inscribed into the analytical technology by design and construction, so that the analysis takes the privacy requirements in consideration from the very start. Hence, privacy will not be seen as a loss of information quality but it may be seen as an enforcement of a right. A complete survey on privacy by design in big data can be found in [19].

From its definition and properties it should be expected that big data anonymization will entail more difficulties and privacy risks than "traditional" anonymization. One of the aims of big data is to increase the information on individuals by gathering and increasingly keeping information from many different sources. This, on one hand, makes it difficult to follow the information path with a corresponding loss of transparency and control; on the other hand combining data sets from different sources, or considering the publication of dynamic databases increases privacy risks. Two publications that involve the same individual may give additional information that put together uniquely identifies her.

The same characteristics of volume, variety, velocity and variability, imply important differences with the "traditional" privacy by design strategies [39] mentioned at the introduction: Minimise, Hide, Separate, Aggregate, Inform, Control, Enforce and Demonstrate.

Big data analytics benefit from gathering greater amounts of information about individuals, most of the time keeping the data and reusing it beyond its original intended use, with the aim of obtaining added value from it. Also, combining different data sources difficults the individual control of his data as he may not even be aware that it is being collected, even less how it travels from one source to another. When there are many data processing entities, informing the users about who has their data and for which purposes is processing it is more complicated. Also opt-in and opt-out mechanisms must be created.

Linkability, composability and computability are requirements that a privacy model must satisfy to be useful for big data anonymization [62]. In this respect, the main limitation of k-anonymity is related to composability since the release of several k-anonymous data sets may lead to re-identification because linking two k-anonymous datasets does not implies that the obtain data set is k-anonymous for any $k > 1$.

On the other hand, differential privacy is composable since linking two ϵ, ϵ'-differentially private datasets has $\epsilon + \epsilon'$-differential privacy. The main shortcoming of differential privacy for big data anonymization is that, in general, differentially private data sets retain not enough analytical utility for exploratory analyses (which are the customary analyses on big data).

The techniques for static databases as they stand do not protect dynamic databases. Streaming anonymization algorithms have been created with such purpose, e.g., [11].

Local anonymization has been proposed as an alternative way to control anonymization by the individuals themselves. Since local anonymization may yield less utility than centralized anonymization, co-utile collaborative anonymization was proposed by [63] which incurs no more information loss than one obtained with the centralized approach and guarantees that neither the data subjects nor the data controller gain more knowledge about the confidential attributes of any other specific data subject than the knowledge contained in the final anonymized data set.

Revealing information about ourselves may reveal information about other people, such as DNA data, our social connections or sharing our locations. Due to the improvements achieved in

machine learning algorithms, personalization has become more precise, extracting from our data a fairly accurate image of who we are based on comparisons to others. While technologies that personalize information and provide people with recommendations can be a valuable assistance, they could also inadvertently or deliberately manipulate people and influence opinions. Psycho-demographic profiles can be extracted at large scale from digital records [43] and could be used in political campaigns to generate messages that appeal to each individual, therefore making them as convincing and relevant as possible. Even sensitive information (such as sexual orientation or religion) can be inferred by apparently unrelated attributes. This could lead to a decrease in trust in online services and at the end it would turn against the industry itself. This could be prevented by providing the users with transparency and control over their information.

However, it has been questioned in [7] if we can no longer rely on control of data to achieve privacy, and in that case propose alternative models to deal with networked privacy, arguing that "we need to let go of our cultural fetishization with the individual as the unit of analysis. We need to develop models that position networks, groups, and communities at the center of our discussion. And we need to find a way to empower people by freeing them to share in ways that don't negatively affect how others lives are interpreted."

The above is in line with the technical perspective that keeping data agreggated up to some point is useful for preserving privacy. By considering that the individual relations are the foundations of the networks and that nowadays all the information is linked, this is also in line with the aim of utility preservation when aggregating data. The big question is: where to put the limit between personalization and privacy?

References

[1] R. Agrawal and R. Srikant. Privacy preserving data mining. In Proc. of the 2000 ACM SIGMOD Conference on Management of Data, pp. 439–450. ACM, (2000).

[2] R. Anderson, R. Needham, and A. Shamir. The steganographic file system. In Information Hiding, volume 1525 of Lecture Notes in Computer Science, pp. 73–82. Springer Berlin, Heidelberg, (1998).

[3] L. Backstrom, C. Dwork, and J. Kleinberg. Where Art Thou R3579X? Anonymized Social Networks, Hidden Patterns, and Structural Steganography. Proc. 16th Intl. World Wide Web Conference (2007).

[4] George R. Blakley. Safeguarding cryptographic keys. In Proc. of the National Computer Conference, 48, pp. 313–317, (1979).

[5] J. Blocki, A. Blum, A. Datta, and O. Sheffet. Differentially private data analysis of social networks via restricted sensitivity. In ITCS, (2013).

[6] D. Boneh, G. di Crescenzo, R. Ostrovsky, and G. Persiano. Public key encryption with keyword search. In Advances in Cryptology EUROCRYPT 04, volume 3027 of Lecture Notes in Computer Science, pp. 506–522. Springer, (2004).

[7] D. Boyd. Networked Privacy. Surveillance & Society, [S.l.], v. 10, n. 3/4, pp. 348–350, dec. (2012). ISSN 1477-7487. Available at: `http://ojs.library.queensu.ca/index.php/surveillance-and-society/article/view/networked`. Date accessed: 30 may 2017.

[8] R. Brand. Microdata protection through noise addition, In J. Domingo-Ferrer, editor, Inference Control in Statistical Databases, volume 2316 of LNCS, pages 97– 116, Berlin Heidelberg, Springer, (2002).

[9] E. F. Brickell and Y. Yacobi. On privacy homomorphisms (extended abstract). In D. Chaum and W. L. Price, editors, EUROCRYPT, volume 304 of Lecture Notes in Computer Science, pp. 117–125. Springer, (1987).

[10] A. Campan and T. M. Truta. A clustering approach for data and structural anonymity in social networks. In Proceedings of the 2nd ACM SIGKDD International Workshop on Privacy, Security, and Trust in KDD (PinKDD'08), in Conjunction with KDD'08, Las Vegas, Nevada, USA, (2008).

[11] J. Cao, B. Carminati, E. Ferrari, and K. Tan. Castle: Continuously anonymizing data streams. IEEE Transactions on Dependable and Secure Computing, 8(3), pp. 337352, (2011).

[12] D. L. Chaum. Untraceable electronic mail, return addresses, and digital pseudonyms. Communications of the ACM, 24(2), pp.84–90, (1981).

[13] S. Chester, B.M. Kapron, G. Ramesh, G. Srivastava, A. Thomo, and S. Venkatesh. Why Waldo befriended the dummy? k-Anonymization of social networks with pseudo-nodes, Soc. Netw. Anal. Min. 3 (3), pp. 381–399 (2013).

[14] B. Chor, O. Goldreich, E. Kushilevitz, and M. Sudan. Private information retrieval. J. ACM, 45(6), pp. 965–981. Nov.(1998).

[15] V. Ciriani, S. De Capitani di Vimercati, S. Foresti, and P. Samarati. Microdata Protection, Secure Data Management in Decentralized Systems, Volume 33 of the series Advances in Information Security, pp. 291–321, (2007).

[16] C. Clifton and T. Tassa. On syntactic anonymity and differential privacy. Transactions on Data Privacy, 6(2), pp. 161-183, (2013).

[17] G. Cormode, C. M. Procopiuc, E. Shen, D. Srivastava, T. Yu. Empirical privacy and empirical utility of anonymized data. In: ICDE Workshop on Privacy-Preserving Data Publication and Analysis (2013).

[18] R. Curtmola, J.A. Garay, S. Kamara, and R. Ostrovsky. Searchable symmetric encryption: Improved definitions and efficient constructions. In A. Juels, R. N. Wright, and S. De Capitani di Vimercati, editors, Conference on Computer and Communications Security (CCS 06). ACM, (2006).

[19] G. D'Acquisto, J. Domingo-Ferrer, P. Kikiras, V. Torra, Y.-A. de Montjoye, and A. Bourka. Privacy by design in big data: An overview of privacy enhancing technologies in the era of big data analytics. CoRR abs/1512.06000, (2015).

[20] T. Dalenius. Towards a methodology for statistical disclosure control. Statistik Tidskrift 15, 429444 (1977).

[21] G. Danezis, J. Domingo-Ferrer, M. Hansen, J-H Hoepman, D. Le Mtayer, R. Tirtea, and S. Schiffner. Privacy and data protection by designfrom policy to engineering. Technical report, ENISA (2015).

[22] J. Domingo-Ferrer, A. Solanas, J. Castellà-Roca. h(k)-private information retrieval from Privacy-Uncooperative Queryable Databases. Online Inf. Rev. 33(4), pp. 720–744, (2009).

[23] J. Domingo-Ferrer and J. Soria-Comas. From t-closeness to differential privacy and vice versa in data anonymization, Knowl.Based Syst. 74, pp. 151–158, (2015).

[24] J. Domingo-Ferrer and V. Torra. Disclosure risk assessment in statistical data protection, Journal of Computational and Applied Mathematics, 164 (1) , pp. 285–293, (2004).

[25] J. Domingo-Ferrer and V. Torra. Ordinal, continuous and heterogeneous k-anonymity through microaggregation. Data Min. Knowl. Discov., 11(2), pp. 195–212, (2005).

[26] J. Domingo-Ferrer and V. Torra. Disclosure protection methods and information loss for microdata. In P. Doyle, J. I. Lane, J.J.M. Theeuwes, and L. Zayatz, editors, Confidentiality, Disclosure and Data Access: Theory and Practical Applications for Statistical Agencies, pp. 91–110, Amsterdam, North-Holland, (2001).

[27] C. Dwork, F. McSherry, K. Nissim, and A. Smith. Calibrating noise to sensitivity in private data analysis. In S. Halevi and T. Rabin, editors, TCC, volume 3876 of Lecture Notes in Computer Science, pp. 265–284. Springer, (2006).

[28] C. Dwork and M. Naor. On the difficulties of disclosure prevention in statistical databases or the case for differential privacy. Journal of Privacy and Confidentiality, 2(1), (2010).

[29] G. T. Duncan, S. A. Keller-McNulty, and S.L. Stokes. Disclosure risk vs. data utility: The R-U condentiality map. Technical report, Los Alamos National Laboratory. LA-UR-01-6428 (2001).

[30] G. T. Duncan and R. W. Pearson. Enhancing access to microdata while protecting confidentiality: prospects for the future. Statistical Science, 6, pp. 219–239, (1991).

[31] T. El-Gamal. A public key cryptosystem and a signature scheme based on discrete logarithms. IEEE Transactions on Information Theory, 31(4), pp. 469-472, (1985).

[32] J. Gehrke, E. Lui, and R. Pass. Towards privacy for social networks: a zero-knowledge based denition of privacy. In: Proceedings of the 8th conference on Theory of cryptography. TCC11, 432–449, (2011).

[33] C. Gentry. Fully homomorphic encryption using ideal lattices. In M. Mitzenmacher, editor, STOC, pp. 169-178. ACM, (2009).

[34] F. Giannotti, D. Pedreschi, S. Pentland, P. Lukowicz, D. Kossmann, J. Crowley, and D.A Helbing. A planetary nervous system for social mining and collective awareness, Eur. Phys. J. Special Topics 214, 49, (2012).

[35] D. Goldschlag, M. Reed, and P. Syverson. Onion routing. Communications of the ACM, 42(2), pp. 39–41, (1999).

[36] M. Hay, C. Li, G. Miklau, and D. Jensen. Accurate estimation of the degree distribution of private networks. In ICDM, (2009).

[37] M. Hay, G. Miklau, D. Jensen, and D. Towsley. Resisting structural identification in anonymized social networks. In Proceedings of the 34th International Conference on Very Large Databases (VLDB'08). ACM, (2008).

[38] H. Hedbom. A survey on transparency tools for enhancing privacy. In V. Matyáš, S. Fischer-Hbner, D. Cvrček, and P. Švenda, editors, The Future of Identity in the Information Society Proc. 4th IFIP WG 9.2, 9.6/11.6, 11.7/FIDIS International Summer School, volume 298 of IFIP Advances in Information and Communication Technology, pp. 67-82. IFIP, Springer, (2009).

[39] J-H. Hoepman. Privacy design strategies (extended abstract). In ICT Systems Security and Privacy Protection - 29th IFIP TC 11 International Conference, SEC 2014, Marrakech, Morocco, June 2-4, 2014. Proceedings, pp. 446-459, (2014).

[40] D.C. Howe and H. Nissenbaum. TrackMeNot: resisting surveillance in web search, in I. Kerr, C. Lucock, and V. Steeves, editors, Lessons From the Identity Trail: Privacy, Anonymity and Identity in a Networked Society, chapter 23. Oxford University Press, (2009).

[41] M. Juárez and V. Torra, Toward a privacy agent for information retrieval. Int. J. Intell. Syst. 28, pp. 606–622 (2013).

[42] S. P. Kasiviswanathan, K. Nissim, S. Raskhodnikova, and A. Smith. Analyzing graphs with node differential privacy. In Theory of Cryptography: 10th Theory of Cryptography Conference, TCC 2013, Tokyo, Japan, March 3-6, (2013).

[43] M. Kosinski, D. Stillwell, and D. Graepel. Private traits and attributes are predictable from digital records of human behavior, PNAS, vol. 110, no. 15, pp. 5802–5805, (2013).

[44] N. Li, T. Li, and S. Venkatasubramanian. t-Closeness: Privacy beyond k-anonymity and ℓ-diversity. In Rada Chirkova, Asuman Dogac, M. Tamer zsu, and Timos K. Sellis, editors, ICDE, pp. 106115. IEEE, (2007).

[45] N. Li, W. H. Qardaji, and D. Su. Provably private data anonymization: Or, k-anonymity meets differential privacy. CoRR, abs/1101.2604, (2011).

[46] N. Li, W. Qardaji, and D. Su. On sampling, anonymization, and differential privacy: Or, k-anonymization meets differential privacy, in 7th ACM Symposium on Information, Computer and Communications Security (ASIACCS2012), Seoul, Korea, May 2-4, (2012).

[47] Y. Lindell and B. Pinkas. Privacy-preserving data mining. In Advances in Cryptology-CRYPTO 2000, volume 1880 of Lecture Notes in Computer Science, pp. 36–54. Springer Berlin / Heidelberg, (2000).

[48] K. Liu and E. Terzi. Towards identity anonymization on graphs. In: Proceedings of the ACM SIGMOD International Conference on Management of Data, pp. 93-106, (2008).

[49] M. Meeker and L. Wu. Internet Trends, (2013).

[50] F. McSherry and K. Talwar. Mechasim Design via Differential Privacy. Proceedings of the 48th Annual Symposium of Foundations of Computer Science, (2007).

[51] P. Paillier. Public-key cryptosystems based on composite degree residuosity classes. In Jacques Stern, editor, EUROCRYPT, volume 1592 of Lecture Notes in Computer Science, pp. 223–238. Springer, (1999).

[52] A. Machanavajjhala, D. Kifer, J. Gehrke, and M. Venkitasubramaniam. l-Diversity: Privacy beyond k-anonymity. ACM Trans. Knowl. Discov. Data, 1(1), (2007).

[53] H. Nissenbaum. Privacy in Context: Technology, Policy, and the Integrity of Social Life. Palo Alto: Stanford Law Books. (2009).

[54] M. O. Rabin. How to exchange secrets with oblivious transfer. Technical Report. TR-81, Aiken Computation Lab, Harvard University, (1981).

[55] R. L. Rivest, L. M. Adleman, and M. L. Dertouzos. On data banks and privacy homomorphisms. In R. A. De Millo et al., editors, Foundations of Secure Computation, pp. 169179, New York, USA. Academic Press, (1978).

[56] R. L. Rivest, A. Shamir, and L. M. Adleman. A method for obtaining digital signatures and public-key cryptosystems. Commun. ACM, 21(2), pp. 120–126, (1978).

[57] J. Salas and V. Torra. Graphic sequences, distances and k-degree anonymity, Disc. Appl. Math. 188, pp. 25–31, (2015).

[58] J. Salas and V. Torra. Improving the characterization of P-stability for applications in network privacy, Disc. Appl. Math. Volume 206, pp. 109–114, (2016).

[59] P. Samarati and L. Sweeney. Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression, Technical Report, SRI International, (1998).

[60] P. Samarati. Protecting respondents identities in microdata release, IEEE Trans. Knowl. Data Eng. 13(6), pp. 1010–1027, (2001).

[61] J. Soria-Comas, J. Domingo-Ferrer, D. Snchez, and S. Martnez. Enhancing data utility in differential privacy via microaggregation-based k-anonymity, The International Journal on Very Large Data Bases (VLDB), vol. 23, no. 5, pp. 771–794, (2014).

[62] J. Soria-Comas and J. Domingo-Ferrer. Big data privacy: challenges to privacy principles and models, Data Science and Engineering, vol. 1, no. 1, pp. 1-8, (2015).

[63] J. Soria-Comas and J. Domingo-Ferrer. Co-utile Collaborative Anonymization of Microdata, in 12th International Conference, MDAI 2015, Skövde, pp. 192–206, (2015).

[64] K. Stokes and V. Torra. Reidentification and k-anonymity: a model for disclosure risk in graphs, Soft Computing. 16(10), pp. 1657–1670, (2012).

[65] L. Sweeney. k-anonymity: a model for protecting privacy, Internat. J. Uncertain. Fuzziness Knowledge-Based Systems 10(5), pp. 557–570, (2002).

[66] A. Shamir. How to share a secret. Communications of the ACM 22 (11), pp. 612613, (1979).

[67] J. van den Hoven, D. Helbing, D. Pedreschi, J. Domingo-Ferrer, F. Gianotti and M. Christen. FuturICT The road towards ethical ICT. EPJ Special Topics 214, pp. 153–181, (2012).

[68] V. S. Verykios and A. Gkoulalas-Divanis. A survey of association rule hiding methods for privacy. In Privacy-Preserving Data Mining: Models and Algorithms, pp. 267–289. Springer, (2008).

[69] A. C. Yao. Protocols for secure computations (extended abstract). In FOCS, pp. 160–164. IEEE Computer Society, (1982).

[70] E. Zheleva and L. Getoor. Preserving the Privacy of Sensitive Relationships in Graph Data. In ACM SIGKDD Workshop on Privacy, Security, and Trust in KDD (PinKDD), pp. 153–171, (2007).

[71] B. Zhou and J. Pei. Preserving privacy in social networks against neighborhood attacks. In ICDE, (2008).

[72] B. Zhou, J. Pei, and W. S. Luk. A brief survey on anonymization techniques for privacy preserving publishing of social network data. ACM SIGKDD Explorations Newsletter, 10 (2), pp. 12–22, (2008).

[73] Personal Data: The Emergence of a New Asset Class. World EconomicForum. `http://www3.weforum.org/docs/WEF_ITTC_PersonalDataNewAsset_Report_2011.pdf`, (2011).

[74] U.S. Dep't. of Health, Education and Welfare, Secretary's Advisory Committee on Automated Personal Data Systems, Records, computers, and the Rights of Citizens viii, (1973).

Julián Salas
Internet Interdisciplinary Institute (IN3)[1], Universitat Oberta de Catalunya (UOC), Barcelona
e-mail: `jsalaspi@uoc.edu`

Josep Domingo-Ferrer
Universitat Rovira i Virgili, Tarragona
e-mail: `josep.domingo@urv.cat`

[1] With the support of a UOC postdoctoral fellowship