TRANSACTIONS ON DATA PRIVACY 16 (2023) 25–49

k-Anonymity: From Theory to Applications

Sabrina De Capitani di Vimercati, Sara Foresti, Giovanni Livraga, Pierangela Samarati

Computer Science Department, Università degli Studi di Milano, Italy

E-mail: {*firstname.lastname*}@unimi.it

Received 22 March 2022; received in revised form 3 August 2022; accepted 3 August 2022

Abstract. k-Anonymity is a well-known privacy model originally designed to protect the identities of the individuals involved in the release of a data collection. It provides a privacy requirement and a metric able to capture the protection degree enjoyed by respondents (i.e., the individuals to whom released data refer). Since its proposal, k-anonymity has been heavily investigated, with works addressing extensions of its privacy requirement to capture specific privacy risks, approaches to efficiently enforce k-anonymity, and adaptations to application scenarios that go beyond the publication of a dataset. In this paper, we illustrate k-anonymity and its main extensions. We also discuss some of the main approaches proposed for the enforcement of the corresponding privacy requirements, and some advanced application scenarios.

Keywords. k-Anonymity, ℓ-Diversity, Privacy, Quasi-identifier, Generalization, Fragmentation, Microaggregation

1 Introduction

k-Anonymity [62] is a well-known privacy model that provides a *privacy requirement* as well as a *privacy metric* able to assess the degree by means of which a data collection to be released satisfies the requirement. k-Anonymity has been originally designed to protect the identities of the individuals to whom the released data items refer (i.e., the data respondents). The privacy requirement pursued by k-anonymity, which enjoys the undeniable benefit of being intuitive and easy to understand, informally requires that each released data item should be indistinguishably related to no less than a certain number of respondents. Such number of respondents is expressed by the value assigned to parameter k, which then can be seen as a metric useful for assessing "how much" the identities of the respondents involved in a release are protected: intuitively, the higher the value of k, the higher the protection enjoyed by respondents. k-Anonymity is typically enforced by producing and releasing, instead of the original dataset, a *sanitized* (i.e., k-anonymous) version of the dataset that satisfies the requirement for the chosen value of k. Sanitization is performed by applying techniques that preserve the truthfulness of the information of each data item. Such techniques include, for example, generalization (publishing more general values for the data) and suppression (removing some data).

While easy to express and understand, k-anonymity needs care in its enforcement to preserve the utility of the released dataset. Intuitively, the more the applied generalization and suppression, the higher the protection: in an extreme scenario, producing and releasing a dataset generalized at the maximum possible level is likely not to pose risks to any individual, but would certainly be of no use for its recipients. This observation highlights the tension between data *privacy* and data *utility*: the less complete a data release is (higher privacy), the less useful it can be for recipients wishing to analyze and use it (lower utility). It is therefore crucial to find the right trade-off between privacy and utility, balancing the need for privacy of data respondents and the need for useful data of data recipients. As it will be illustrated in this paper, computing *optimal* k-anonymous datasets that respect the k-anonymity requirement without over-protecting the dataset and maintaining utility of the sanitized data is computationally hard, and several approaches have been proposed in this regard.

Since its introduction, k-anonymity has been extensively studied for proposing and enforcing extended/revised privacy requirements suited, for example, to specific privacy risks and/or release scenarios, also with practical impact on privacy regulations. k-Anonymity offers a starting point for protecting data privacy and clearly by itself is not sufficient to tackle the complex privacy problems. It has then be a subject of discussions concerning the underlying idea, pursued also by its extensions, of protecting respondents by producing sanitized versions of datasets. Differential privacy, for example, takes a different approach and perturbs the results of analyses over unprotected datasets [32]. In particular, the definition of differential privacy and its extensions demand that the output of an analysis should not depend 'too much' on the involvement in the computation of the data of any specific individual. Typically, algorithms that satisfy the differential privacy requirement add random noise to the result of their computations to guarantee that, for any given pair of neighbor datasets D and D' (i.e., datasets that differ in only one record), the probability of observing a result on D is close to that of observing the same result on D'. Such closeness depends on a parameter ϵ called privacy budget. Recent research has showed that both approaches based on k-anonymity and approaches based on differential privacy have their pros and cons [6, 18, 19, 60], and could also be jointly adopted to ensure protection while producing useful results (e.g., [50, 66]). Also, k-anonymity has recently been adopted, possibly enforcing slightly reformulated privacy requirements, in different application scenarios that depart from the publication of datasets. Among these approaches, it is worth noting a recent solution for checking compromised credentials that permits a user to privately check whether her passwords had been compromised and are part of a database of stolen passwords without releasing the precise passwords to be checked [47].

In this paper, we illustrate the privacy models and requirements of k-anonymity and of some of its extensions, the main approaches proposed for the enforcement of such requirements, and some advanced application scenarios. The remainder of this paper is organized as follows. In Section 2, we discuss the theory and main privacy requirements pursued by k-anonymity and some of its well-known extensions. In Section 3, we focus on the enforcement of these privacy requirements, illustrating solutions based on data generalization, data fragmentation, and microaggregation. In Section 4, we turn our attention to advanced application scenarios that have seen the application of k-anonymity for protecting user privacy, including location-based services and the publication of movement data, analysis of social network data, contact tracing applications, and big data analytics. In Section 5, we provide our conclusions.

SSN	LastName	FirstName	DoB	Sex	ZIP	Disease
123-45-6789	Alice	Ant	1940/08/10	F	98512	Heart attack
234-56-7890	Bob	Bell	1950/02/12	M	99413	COVID-19
345-67-8901	Carol	Candle	1940/08/04	F	98578	Cardiomyopathy
456-78-9012	David	Dart	1950/02/13	M	99356	COVID-19
567-89-0123	Eric	Eel	1950/07/12	M	99423	Dermatitis
678-90-1234	Fred	Frog	1940/08/11	F	98545	Pericarditis
789-01-2345	Greg	Glace	1950/07/25	M	99334	Short breath
890-12-3456	Hal	Heart	1950/07/30	M	99490	Cough
901-23-4567	Ian	Instance	1950/02/20	M	99301	COVID-19
012-34-5678	Luke	Lane	1945/12/01	M	98321	Astrocytoma

Figure 1: An example of a dataset including personal and medical information for a set of patients

2 Theory and Privacy Requirements

In this section, we illustrate the anonymity problem originally addressed by k-anonymity (Section 2.1) and some of the most well-known privacy requirements (Sections 2.2 and 2.3) that may be enforced when a dataset needs to be released and/or shared.

2.1 The Anonymity Problem

k-Anonymity and its variations have been originally designed to operate in scenarios where the datasets to be released or shared are represented as *microdata tables* defined over a set of attributes of interest, with a record for each possible respondent. Based on their characteristics, attributes composing a dataset can be classified in: *i) identifying* attributes that can *per se* univocally identify the respondent of a record (e.g., SSN or name); *ii) quasi-identifying* (QI) attributes that can be *linked to external data sources* to re-identify the respondent of a record (e.g., date of birth, sex, address); *iii) sensitive* attributes that represent respondents' sensitive information (e.g., disease, income); and *iv) non-sensitive* attributes that are unlikely to permit re-identification and are not sensitive. Figure 1 illustrates an example of a fictitious medical dataset, defined over a set of attributes representing personal information (`SSN`, `LastName`, `FirstName`, `DoB`, `Sex`, and `ZIP`) and medical (`Disease`) information for 10 respondents.

The first step for protecting the privacy of the respondents involved in a data release requires to *de-identify* respondents, hiding (e.g., by removing or encrypting) identifying attributes. Figure 2(a) illustrates a de-identified version of the medical dataset in Figure 1, where identifiers `SSN`, `LastName`, and `FirstName` are removed. Unfortunately, de-identification does not provide any guarantee of *anonymity*. This is due to the presence of quasi-identifying attributes (in our example, attributes `DoB`, `Sex`, and `ZIP`), which may be linked to external, non de-identified information to re-identify (some of) the respondents. Figure 2(b) illustrates an excerpt of a (fictitious) voter list, which could be linked to the medical dataset through QI attributes `DoB`, `Sex`, and `ZIP`. In particular, the medical dataset contains only a record for a male respondent, born on December 1st, 1945, and living in the 98321 area. If this combination is unique in the external world as well, it permits recipients to re-identify the last record as pertaining to Luke Lane (*identity disclosure*), disclosing also the fact that his diagnosis is astrocytoma (*attribute disclosure*).

SSN	LastName	FirstName	DoB	Sex	ZIP	Disease
			1940/08/10	F	98512	Heart attack
			1950/02/12	M	99413	COVID-19
			1940/08/04	F	98578	Cardiomyopathy
			1950/02/13	M	99356	COVID-19
			1950/07/12	M	99423	Dermatitis
			1940/08/11	F	98545	Pericarditis
			1950/07/25	M	99334	Short breath
			1950/07/30	M	99490	Cough
			1950/02/20	M	99301	COVID-19
			1945/12/01	*M*	*98321*	*Astrocytoma*

(a)

Name	Address	City	ZIP	DoB	Sex
...	...	...	...	...	...
Luke Lane	10 Cedar St.	Buckley, WA	*98321*	*45/12/01*	*male*
...	...	...	...	...	...

(b)

Figure 2: A de-identified version of the dataset in Figure 1 (a) and an example of a publicly available non de-identified dataset (b)

Removing –besides identifiers– also quasi-identifying attributes to prevent the improper disclosure illustrated above is clearly not a feasible approach, since such attributes typically represent a large portion of the informative content of a dataset and their complete removal could make the dataset useless for recipients. In the remainder of this section, we discuss some of the most well-known privacy requirements that may be adopted to effectively protect the privacy of the respondents involved in a data release.

2.2 k-Anonymity

The privacy requirement enforced by k-anonymity [62] demands that *any released information should be indistinguishably related to no less than a certain number (k) of respondents*. In the context of the re-identification illustrated above, this privacy requirement has been translated by Samarati [62] into the k-anonymity requirement, which requires each release of data to be such that *every combination of quasi-identifying values can be indistinctly matched to at least k respondents*. Following the typical assumption that each respondent is represented by a single record in the dataset to be released, the requirement of k-anonymity is satisfied by a dataset if each record in the dataset cannot be related to less than k individuals in the population and, conversely, each individual in the population cannot be related to less than k records in the dataset.

The verification of the k-anonymity requirement would require knowledge of the external data sources that may be used for linking through quasi-identifying attributes. This assumption is unfeasible in practice, and hence a safe approach is to consider a dataset k-anonymous if each combination of quasi-identifying values appearing in the dataset has at least k occurrences. In this way, each respondent in a k-anonymous dataset is indistinguishable (w.r.t. QI attributes) from at least $k-1$ other respondents in the same dataset [26]. For example, the dataset in Figure 2(a) is 1-anonymous considering QI=$\{$DoB, Sex, ZIP$\}$, as it in-

SSN	LastName	FirstName	DoB	Sex	ZIP	Disease
			1940/08/**	F	98***	Heart attack
			1940/08/**	F	98***	Cardiomyopathy
			1940/08/**	F	98***	Pericarditis
			1950/02/**	M	99***	COVID-19
			1950/02/**	M	99***	COVID-19
			1950/02/**	M	99***	COVID-19
			1950/07/**	M	99***	Dermatitis
			1950/07/**	M	99***	Short breath
			1950/07/**	M	99***	Cough

Figure 3: An example of a 3-anonymous version of the dataset in Figure 2(a)

cludes unique combinations of quasi-identifying values (e.g., $\langle$1940/08/10,F,98512$\rangle$). The definition above of k-anonymous datasets represents a sufficient condition to satisfy the k-anonymity requirement: since each quasi-identifying value appears in a k-anonymous dataset with at least k occurrences, it is immediate to see that each respondent can be matched to no less than k records in the dataset.

Different k-anonymous versions of a dataset may be obtained in different ways (Section 3). The first and most studied approach in this line of research modifies the quasi-identifying attributes only, replacing their values with more general values so that each value appears with at least k occurrences. Replacing values with more general ones is a data protection technique called *generalization* [33]. Generalization is typically applied together with *suppression*, another data protection technique that consists in selectively removing data items from a dataset (typically, in the context of k-anonymity, suppression is adopted to remove few outliers that would force a large amount of generalization for k-anonymity satisfaction). Figure 3 represents a 3-anonymous version of the de-identified medical dataset in Figure 2(a), where quasi-identifying attributes `DoB` and `ZIP` have been generalized, and the record of Luke Lane (last record in Figure 2(a)) has been suppressed. In particular, `DoB` has been generalized by releasing only the year and month of birth (hiding the day of the month), and `ZIP` has been generalized by releasing only the first two digits. Note that maintaining Luke Lane's record would have required more generalization to achieve 3-anonymity, since the same generalization would not have made his quasi-identifying values equal to those of at least 2 other records (e.g., its generalized `DoB` value would be 1945/12/**, appearing with one occurrence only in the dataset).

Generalization and suppression enjoy the undeniable benefits of producing k-anonymous *truthful* datasets, as no data item is perturbed in its values (which are simply made more general or suppressed). However, protection through generalization and suppression comes at the price of reducing details from the dataset to be released. To the aim of minimizing such information loss, it is necessary to compute k-anonymous datasets while minimizing the adoption of generalization and suppression. To this end, both exact and heuristic algorithms have been proposed [16]. Note that, as we will illustrate in Section 3, there are proposals that achieve k-anonymity without resorting to generalization and suppression, adopting other data protection techniques (and possibly producing non-truthful information).

SSN	LastName	FirstName	DoB	Sex	ZIP	Disease
			1940/**/**	F	985**	Heart attack
			1940/**/**	F	985**	Cardiomyopathy
			1940/**/**	F	985**	Pericarditis
			1950/**/**	M	994**	COVID-19
			1950/**/**	M	994**	Dermatitis
			1950/**/**	M	994**	Cough
			1950/**/**	M	993**	Short breath
			1950/**/**	M	993**	COVID-19
			1950/**/**	M	993**	COVID-19

Figure 4: An example of a 3-anonymous and 2-diverse version of the dataset in Figure 2(a)

2.3 ℓ-Diversity, t-Closeness and Extensions

The privacy requirement of k-anonymity is explicitly formulated to protect respondents' identities, is only a starting point for protecting privacy and further steps are needed to prevent attribute disclosure, which can be otherwise possible due to value *homogeneity* and *external knowledge* attacks, as follows.

- The problem of *homogeneity* occurs when all records in an *equivalence class* (i.e., the set of at least k records with the same value for QI attributes) in a k-anonymous dataset assume the same value for the sensitive attribute.
- The *external knowledge attack* occurs when a data recipient can successfully leverage external knowledge she possesses on a respondent to reduce the uncertainty about the sensitive attribute value of that respondent.

To illustrate these problems, consider the 3-anonymous dataset in Figure 3. A recipient who knows that a male individual, born on 1950/02/12 and living in the 99413 area, is included in the dataset can immediately learn that he is represented in the second equivalence class. Even without pinpointing his specific record, by accessing the 3-anonymous dataset the recipient can discover that he suffers from COVID-19, which is the only value assumed by all records in the class (homogeneity attack). Assume now that a recipient knows that a male friend of hers, born on 1950/07/12 and living in 99423 area, is included in the dataset and is also practicing for a marathon. By accessing the 3-anonymous dataset, the recipient can learn that his friend is represented in the last equivalence class. Due to the fact that short breath and cough would not be possible for a runner practicing for a marathon, the recipient can discover that her friend suffers from dermatitis (external knowledge). To counteract these attacks and limit the risk of attribute disclosure, Machanavajjhala et al. [53] propose the ℓ-diversity requirement, which extends the k-anonymity requirement by demanding that each equivalence class contains at least ℓ *well-represented* values for the sensitive attribute. Several definitions for 'well-represented' values have been proposed, and an intuitive interpretation considers ℓ values well-represented if they are different. Figure 4 illustrates an example of a 3-anonymous and 2-diverse version of the dataset in Figure 2(a) (again after the removal of the last outlier record pertaining to Luke Lane) where each equivalence class contains at least 2 values for sensitive attribute `Disease`.

Like for k-anonymity, the problem of computing ℓ-diverse datasets minimizing information loss due to generalization and suppression is computationally hard. It is interesting

to note that the original algorithms proposed to compute a k-anonymous dataset can be adapted to guarantee also ℓ-diversity by simply considering the diversity of the sensitive values in the equivalence classes.

Li et al. [49] identify two further attacks that may cause improper leakage of sensitive values from a k-anonymous and ℓ-diverse dataset, as follows.

- The *skewness attack* occurs when the value distribution for the sensitive attribute observable in an ℓ-diverse equivalence class differs from that observable in the overall population (or in the overall ℓ-diverse dataset).
- The *similarity attack* occurs when the sensitive values in an ℓ-diverse equivalence class, despite different, are semantically similar.

To illustrate these attacks, consider the 2-diverse dataset in Figure 4. Two records out of three in the last equivalence class assume value COVID-19, meaning that respondents in this equivalence class have over 0.6 probability to suffer from COVID-19. If the probability of suffering from COVID-19 in the population is significantly lower than that, the identification of this equivalence class as the one that includes a respondent signals an increase in the probability that such respondent suffers from COVID-19 (skewness attack). A recipient who knows that a female individual is included in the dataset can immediately learn that she is represented in the first equivalence class, and therefore learn that she suffers from a cardiovascular disease because all values for attribute `Disease` are cardiovascular diseases (similarity attack). To counteract these attacks, Li et al. [49] propose the t-closeness requirement, which extends the k-anonymity requirement by demanding that the value distribution for the sensitive attribute in each equivalence class is similar (i.e., with distance smaller than a threshold t) to that observable in the overall released dataset. It is easy to see that the satisfaction of such a requirement limits the effectiveness of both the similarity and the skewness attack: the presence of semantically similar values in an equivalence class is due to the presence of the same values in the overall dataset with comparable frequencies (reducing the effectiveness of the similarity attack); and no major differences among value distributions can be leveraged to infer sensitive information (reducing the effectiveness of the skewness attack).

ℓ-Diversity and t-closeness represent two well-known extensions of k-anonymity, counteracting specific attacks through imposing more complex privacy requirements to be enforced. Further extensions have been proposed over the years, suited to more complex data release scenarios including the sanitized release of, for example, datasets with multiple records per respondents (e.g., [69, 76]), multiple datasets at the same time or different versions of the same dataset over the time (e.g., [29, 58, 61, 63, 76, 82]), streams of data (e.g., [29, 46, 77, 86]), datasets where the quasi-identifier may be represented by information other than a set of attributes (e.g., [69]), non-relational datasets such as RDF graphs or documents (e.g., [20, 70]). Other extensions have investigated, among other aspects, the possibility of supporting different privacy requirements for different respondents (e.g., [34, 81]) and limiting the leakage of specific values for sensitive attributes (e.g., [79]). Other extensions have considered enforcing k-anonymity in scenarios where any attribute can be considered as a quasi-identifier or as a sensitive attribute (e.g., for anonymizing set-valued data [69]).

3 k-Anonymity Enforcement

In this section, we illustrate how k-anonymity can be efficiently enforced in practice. We discuss some of the most studied families of approaches, based on data generalization (Section 3.1) and on data fragmentation and microaggregation (Section 3.2).

3.1 Generalization-Based Enforcement

As illustrated in Section 2, the first strategy investigated for enforcing k-anonymity leverages generalization, possibly coupled with suppression of outlier data. In principle, both generalization and suppression can be applied at different granularity levels: generalization can be applied at the *cell* level (substituting a cell value with a more general value) or at the *attribute* level (generalizing all the cells in the column), while suppression can be applied at the *cell*, *attribute*, or *record* level (removing a single cell, a column, or a row, respectively) [16]. Regardless of the granularity level at which generalization (and suppression) operates, two main families of approaches have been investigated, which differ in how generalization is defined and operates.

- *Hierarchy-based* generalization leverages the definition of a pre-defined *generalization hierarchy* for each of the quasi-identifying attributes, where the maximal element is the most general value that can be defined for the attribute, and the minimal elements are the most specific values (i.e., those appearing in the original dataset).
- *Recoding-based* generalization generalizes the values of the quasi-identifying attributes into intervals (recoding), typically at runtime and hence does not require the definition of pre-defined hierarchies.

Several approaches have been proposed for enforcing k-anonymity adopting both hierarchy-based and recoding-based generalization [17]. In the reminder of this section, we illustrate some of the most well-known enforcement approaches according to this categorization.

Hierarchy-based generalization (e.g., [44, 62]). Approaches based on hierarchy-based generalization need a formal representation of how generalization operates on attribute values. To formally reason about generalizations, the notion of attribute domain (i.e., the set of values that can be assumed by an attribute of a dataset) is extended to the notion of *generalized domain*: a generalized domain for attribute a contains generalized values that can be used when generalizing a. Different generalized domains can be defined for an attribute, which include the generalized values obtained through the application of different (less or more) generalization steps. For example, attribute `ZIP` could be generalized removing the last digit (one generalization step), the last two digits (two generalization steps), and so on. Given the set of (original and generalized) domains, a *generalization relationship* $\leq_D$ can be defined such that $D_i \leq_D D_j$, with D_i and D_j two domains, iff the values in D_j are generalizations of the values in D_i. In particular, a generalization relationship $\leq_D$ must guarantee that: *i)* the set of domain generalizations for D_i is totally ordered to guarantee determinism in the generalization process; and *ii)* all values in a domain can be generalized to a single value. The definition of the generalization relationship $\leq_D$ implies the existence of a totally ordered hierarchy DGH_D for each domain D, called *domain generalization hierarchy*, which can be graphically represented as a chain of vertices in which the top element is the singleton most generalized domain, and the bottom element is D. Figure 5 illustrates

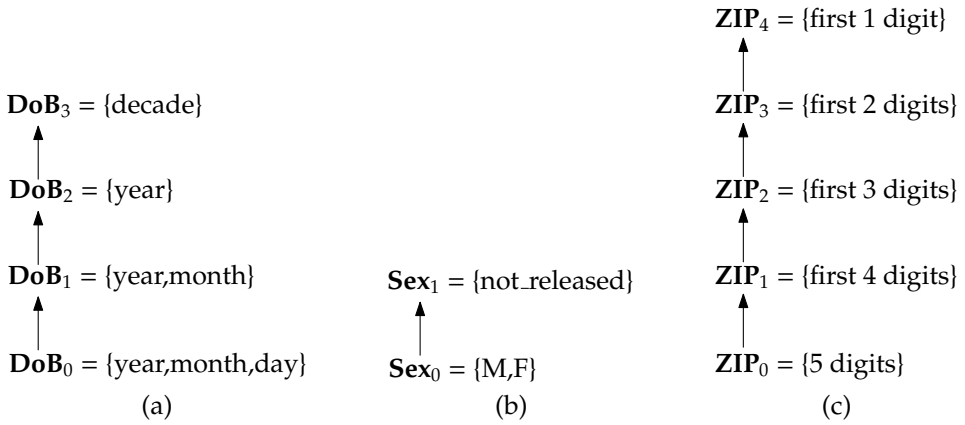


Figure 5: An example of domain generalization hierarchies for domains $\mathbf{DoB}_0$ (a), $\mathbf{Sex}_0$ (b), and $\mathbf{ZIP}_0$ (c)

three examples of domain generalization hierarchies for the quasi-identifying attributes DoB, Sex, and ZIP of the dataset in Figure 2(a). For example, attribute DoB, whose original domain $\mathbf{DoB}_0$ assumes values of the form ⟨year,month,day⟩ (e.g., 1940/08/10 for the first record in Figure 2(a)), can be generalized hiding the day of birth (generalized domain $\mathbf{DoB}_1$), the day and month (generalized domain $\mathbf{DoB}_2$), and by clustering years in decades (generalized domain $\mathbf{DoB}_3$, at which point all records in Figure 2(a) would be generalized to the same value). Similarly, attribute Sex, whose original domain $\mathbf{Sex}_0$ only includes values M and F, can be generalized to value *not_released* ($\mathbf{Sex}_1$). Attribute ZIP, whose original domain $\mathbf{ZIP}_0$ contains 5-digit codes, can be generalized by removing at each step the less significative digit left at the previous generalization step.

Since typically the quasi-identifier of a dataset is composed of a set of attributes that should be generalized, the domain generalization hierarchy definition can be extended to *tuples of domains*: a domain tuple $DT = \langle D_1, \ldots, D_n \rangle$ is an *ordered* set of domains, composed through the Cartesian product. Since the domain generalization hierarchy DGH_{D_i} of each domain D_i in DT is totally ordered, the domain generalization hierarchy DGH_{DT} of domain tuple $DT = \langle D_i, \ldots, D_n \rangle$, defined as $\mathsf{DGH}_{DT} = \mathsf{DGH}_{D_i} \times \ldots \times \mathsf{DGH}_{D_n}$, is a *lattice* where the minimal element is DT and the maximal element is the tuple composed of all top elements in $\mathsf{DGH}_{D_i}, i = 1, \ldots, n$. Figure 6 illustrates a domain generalization hierarchy for domain tuple $\langle \mathbf{DoB}_0, \mathbf{Sex}_0 \rangle$ (i.e., considering the pair of attributes ⟨DoB, Sex⟩) obtained combining the domain generalization hierarchies in Figures 5(a)–(b). Intuitively, each element in the hierarchy corresponds to a generalized dataset where the quasi-identifying attributes are generalized according to the tuple represented by the element. For example, consider the dataset in Figure 2(a), and suppose that DoB and Sex are the quasi-identifying attributes. The 3-anonymous dataset in Figure 3 corresponds to element $\langle \mathbf{DoB}_1, \mathbf{ZIP}_3 \rangle$ of the lattice for domain tuple $\langle \mathbf{DoB}_0, \mathbf{ZIP}_0 \rangle$, while the 2-diverse dataset in Figure 4 corresponds to element $\langle \mathbf{DoB}_2, \mathbf{ZIP}_2 \rangle$. Each path in DGH_{DT} from DT to the maximal element of DGH_{DT} defines a *generalization strategy* that can be adopted when generalizing a quasi-identifier composed of attributes $a_1, \ldots, a_n$ defined over domains $D_1, \ldots, D_n$.

The problem of computing a k-anonymous dataset minimizing generalization (and suppression) and using generalization hierarchies has been proved to be NP-hard [17] and, for this reason, both exact and heuristic approaches have been defined. Two of the most well-

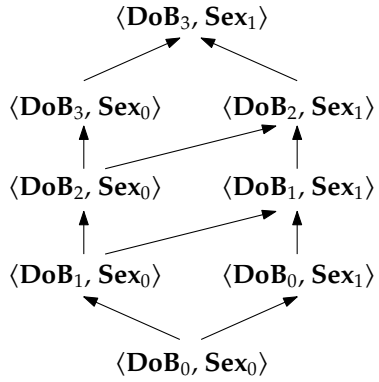


Figure 6: An example of domain generalization hierarchy for domain tuple ⟨**DoB**$_0$,**Sex**$_0$⟩

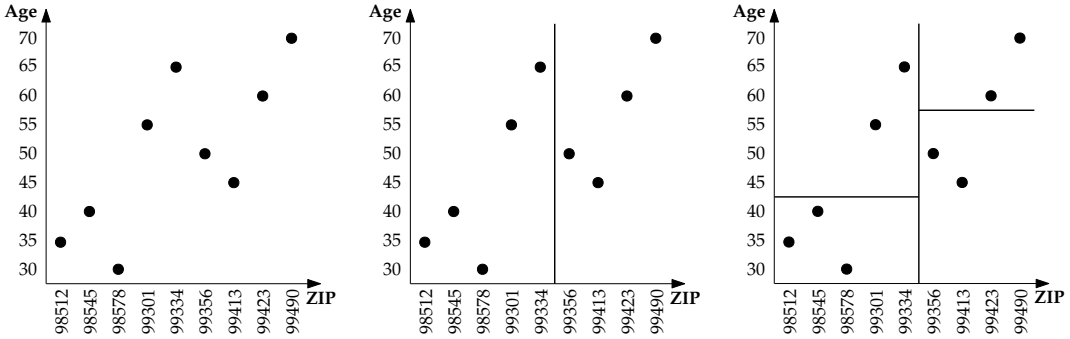
known (exact) approaches in this context are the one proposed in the original k-anonymity proposal by Samarati [62] and the one by LeFevre et al. [44]. In particular, Samarati [62] proposes a solution that leverages binary search on the domain generalization hierarchy to avoid searching in the whole generalization space. The approach is based on the specification of a suppression threshold MaxSup representing the maximum number of suppressed records that is considered acceptable, and on the observation that going up in the hierarchy the number of records that should be suppressed to satisfy k-anonymity decreases: if at a level l of the hierarchy there is no element that corresponds to a k-anonymous dataset suppressing MaxSup or less than MaxSup records, then there cannot be at level l' lower than l in the hierarchy. A binary search is adopted to determine the lowest level at which there is an element that corresponds to a k-anonymous dataset respecting the MaxSup constraint. While permitting to avoid a complete search in the generalization hierarchy, this approach would still require to compute the datasets to check for the k-anonymity requirement. To avoid this, the approach in [62] builds on the notion of (lattice of) *distance vectors* representing distances among the generalized domains, by means of which it is possible to check the satisfaction of k-anonymity without computing the actual corresponding datasets.

LeFevre et at. [44] leverage the observation that if a dataset is k-anonymous considering a set QI of quasi-identifying attributes, then it is also k-anonymous considering any subset of QI as the quasi-identifying attributes. In other words, k-anonymity w.r.t. a set QI$'\subset$QI of quasi-identifying attributes is a necessary (but not sufficient) condition for satisfying k-anonymity w.r.t. QI. The solution in [44] leverages this observation and proposes an approach, known as *Incognito*, which efficiently computes k-anonymous datasets minimizing the adoption of generalization and suppression. Incognito adopts an iterative process, considering at the i^{th} iteration a number i of quasi-identifying attributes, and terminates at iteration $|$QI$|$. To illustrate, consider a dataset with QI$=\{a_1, \ldots, a_n\}$. At the first iteration (i=1) Incognito checks k-anonymity considering each attribute $a_1, \ldots, a_n$ in QI in isolation: the generalizations that do not satisfy k-anonymity are discarded. At the second iteration (i=2), Incognito combines in pairs the generalizations that survived at the first iteration, and checks the satisfaction of the k-anonymity w.r.t. pairs of quasi-identifying attributes (note that all combinations of two quasi-identifying attributes are evaluated). The process continues until all attributes in QI are evaluated (i.e., $i=|$QI$|$). Note that at each step, for each combination of quasi-identifying attributes, when a generalization satisfies k-anonymity, also its direct generalizations do so and are no more evaluated.

Recoding-based generalization (e.g., [9, 45]). A different family of enforcement approaches, which do not require the availability of pre-defined generalization hierarchies, computes –typically at runtime– the intervals (or, more generally, sets) for recoding. These approaches require an ordering among the values of the quasi-identifying attributes. A well-known heuristic approach along this line of research for enforcing k-anonymity, generalizing at the granularity level of single cells, is Mondrian [45]. Mondrian leverages a spatial representation of the data, with QI attributes as dimensions, and (multiset) combinations of QI values appearing in the original dataset as (multisets of) points in this space. Mondrian operates a recursive process to partition the space in regions containing a certain number of points (which corresponds to splitting the original dataset in fragments that contain a certain number of records). In particular, at each iteration, a quasi-identifying attribute a is selected and the regions (corresponding to fragments) obtained at the previous iteration (the entire space, at the first iteration) are split based on the values of a. For example, if a is numerical, regions are split in two sub-regions, one including points whose value for a is lower than or equal to the median, and the other including points with a value for a higher than the median (if a is not numerical, a possibility is to use the value in the median position in the ordering defined over a's values). Such recursive partitioning terminates when any further partitioning would generate a region with less than k points (fragments with less than k records). The values of the quasi-identifying attributes in the resulting regions (fragments) are substituted with a recoding-based generalization. To illustrate, consider the dataset in Figure 7(a), where attributes `Age` and `ZIP` are considered quasi-identifying (note that we consider only two attributes for the QI to ensure clarity of the graphical representation of Mondrian), and suppose that we have to compute a 2-anonymous version of the dataset. Figure 7(b) illustrates the graphical representation of the dataset in Figure 7(a): the space has two dimensions, one for each QI attribute, and points in the space represent the combinations of QI values in the records of the dataset. The space is then recursively partitioned by Mondrian. Figure 7(c) illustrates the first partitioning performed on the dimension representing attribute `ZIP`: the `ZIP` value in median position is 99334, and the partitioning places all points (records) with a `ZIP` value lower than or equal to 99334 in a region (fragment) $R_{\texttt{ZIP}\leq 99334}$, and the remaining ones in another region $R_{\texttt{ZIP}>99334}$. Figure 7(d) illustrates the subsequent partitioning performed on the dimension representing attribute `Age`. The two regions obtained in the first partitioning are further split into two regions each. In particular, in $R_{\texttt{ZIP}\leq 99334}$ the median value for `Age` is 40, and the partitioning places all points (records) in $R_{\texttt{ZIP}\leq 99334}$ that have an `Age` value lower than or equal to 40 in a region (fragment) $R_{(\texttt{ZIP}\leq 99334)\wedge(\texttt{Age}\leq 40)}$, and the remaining ones in another region $R_{(\texttt{ZIP}\leq 99334)\wedge(\texttt{Age}>40)}$. A similar partitioning is done for $R_{\texttt{ZIP}>99334}$, where the median value for `Age` is 55 and hence the partitioning places all points in $R_{\texttt{ZIP}>99334}$ that have an `Age` value lower than or equal to 55 in a region $R_{(\texttt{ZIP}>99334)\wedge(\texttt{Age}\leq 55)}$, and the remaining ones in region $R_{(\texttt{ZIP}>99334)\wedge(\texttt{Age}>55)}$. At this point, no further partitioning is possible without producing regions with less than k=2 points (all regions have already two points but one, which has three points and cannot be split further). The recursion terminates, and the records represented by points in each region are generalized to a same value, thus satisfying 2-anonymity and obtaining the dataset in Figure 7(e). In this example, generalization has been obtained by recoding attribute `Age` into intervals and attribute `ZIP` into sets enumerating their elements. We close this illustration of Mondrian by noting that the approach has been recently extended and adopted for operating in parallel leveraging the computational power of a set of workers and to also enforce the ℓ-diversity requirement [22, 23].

Age	ZIP	Disease
35	98512	Heart attack
45	99413	COVID-19
30	98578	Cardiomyopathy
50	99356	COVID-19
60	99423	Dermatisis
40	98545	Pericarditis
65	99334	Short breath
70	99490	Cough
55	99301	COVID-19

(a) Original dataset



(b) Spatial representation (c) Partitioning over ZIP (d) Partitioning over Age

Age	ZIP	Disease
[30,40]	{98512, 98578,98545}	Heart attack
[30,40]	{98512, 98578,98545}	Cardiomyopathy
[30,40]	{98512, 98578,98545}	Pericarditis
[55,65]	{99301, 99334}	COVID-19
[55,65]	{99301, 99334}	Short breath
[45,50]	{99356, 99413}	COVID-19
[45,50]	{99356, 99413}	COVID-19
[60,70]	{99423, 99490}	Dermatisis
[60,70]	{99423, 99490}	Cough

(e) Recoding-based 2-anonymous dataset

Figure 7: Computation of a 2-anonymous dataset with the Mondrian approach: original dataset with QI={Age,ZIP} (a) and its spatial representation (b), partitioning process (c)–(d), and 2-anonymous version of the dataset (e)

3.2 Fragmentation- and Microaggregation-Based Enforcement

The proposals discussed so far assume to enforce k-anonymity by generalizing the QI attributes (with the possible suppression of outliers). Two different lines of work have investigated the possibility of adopting data fragmentation and microaggregation.

Data fragmentation. An alternative approach to generalization (and suppression) for pro-

SSN	LastName	FirstName	DoB	Sex	ZIP	ID
			1940/08/10	F	98512	
			1940/08/04	F	98578	1
			1940/08/11	F	98545	
			1950/02/12	M	99413	
			1950/07/12	M	99423	2
			1950/07/30	M	99490	
			1950/02/13	M	99356	
			1950/07/25	M	99334	3
			1950/02/20	M	99301	

ID	Disease	Count
1	Heart attack	1
	Cardiomyopathy	1
	Pericarditis	1
2	COVID-19	1
	Dermatisis	1
	Cough	1
3	Short breath	1
	COVID-19	2

Figure 8: An example of a 3-anonymous and 2-diverse version of the dataset in Figure 2(a) obtained with the Anatomy approach

tecting respondents, while still producing truthful information, consists in vertically splitting the original dataset in fragments, for example, to hide the precise correspondence between quasi-identifying and sensitive information and release it at the coarser (and hence more protected) granularity level of groups of records. Following this intuition, Xiao and Tao [80] propose a fragmentation-based approach, called *Anatomy*, to produce ℓ-diverse datasets. Anatomy operates by first grouping the original de-identified records in groups that contain at least ℓ well-represented sensitive values. These groups are then split so to separate quasi-identifying attributes from the sensitive attributes. Figure 8 illustrates a 3-anonymous and 2-diverse version of the dataset in Figure 2(a), again after the suppression of the last record of Luke Lane. The two fragments are complemented with an attribute `ID`, which can be used to link the sub-records in the two fragments at the granularity level of group. The fragment including the sensitive attribute also has an additional attribute `Count`, which associates with each sensitive value the number of occurrences in the specific group (e.g., in the third group value COVID-19 appears twice). It is easy to see that the dataset in Figure 8 guarantees the same degree of protection as the dataset in Figure 4, since any combination of QI values of respondents can be matched to at least 2 different values for the sensitive attribute `Disease`. It is interesting to note that a similar approach can be successfully adopted also whenever the release of a dataset requires to protect and hide *generic* sensitive associations among data (e.g., [3, 21, 24, 25]).

Microaggregation. Another alternative approach for achieving k-anonymity is based on *microaggregation* (e.g., [29, 30, 64, 66]), a perturbative (unlike generalization and fragmentation) data protection technique, which can be operationally defined in terms of a *partitioning* step, followed by an *aggregation* step. In the partitioning step, similarly to what is done in fragmentation-based approaches, the set of records composing the original dataset is partitioned into different clusters in such a way that *i)* records in the same cluster are similar to each other (e.g., for numerical attributes, clusters contain close values), and *ii)* each cluster contains at least k records. In the aggregation step, an aggregation operator (e.g., the mean for continuous data or the median for categorical data) is computed for each cluster and over each QI attribute: the computed result replaces the original values of the attribute on which it has been computed. Since each cluster contains at least k records, each combination of quasi-identifying values appears at least k times. For instance, consider the dataset in Figure 9(a), which –for simplicity– is a view of the dataset in Figure 7(a) over attributes `Age` and `Disease`, where `Age` is a QI attribute and `Disease` is sensitive.

Age	Disease
35	Heart attack
45	COVID-19
30	Cardiomyopathy
50	COVID-19
60	Dermatisis
40	Pericarditis
65	Short breath
70	Cough
55	COVID-19

(a)

Age	Disease
35	Cardiomyopathy
35	Heart attack
35	Pericarditis
50	COVID-19
50	COVID-19
50	COVID-19
65	Dermatisis
65	Short breath
65	Cough

(b)

Figure 9: An example of a dataset (a) and of a 3-anonymous version of it (b) obtained adopting microaggregation and assuming QI={Age}

Figure 9(b) illustrates a 3-anonymous version of the dataset in Figure 9(a) obtained through microaggregation: the records have been clustered in three groups according to a similarity in the values of attribute Age (in this example, according to an ordering over them), and the quasi-identifying values in each group are replaced with the mean. Being microaggregation a perturbative protection technique, k-anonymous datasets computed adopting this approach do not preserve data truthfulness (e.g., all records in Figure 9(b) but the second, fifth and eighth are not real records according to the original values in Figure 9(a)).

It is worth noticing that microaggregation besides being enforced *per se* as illustrated above, can also be used to improve the utility of differentially private responses to arbitrary queries against a dataset [66]. As mentioned in the introduction, given a query to be evaluated on a dataset, a differentially private algorithm adds random noise to the actual result to ensure that the returned result does not strongly depend on the specific data of any individual in the dataset. Clearly, the larger the amount of noise, the lesser the utility for the recipients of the results. In [66], Soria-Comas et al. showed that the amount of noise to be added to query results can be reduced by running the query of interest on a microaggregated k-anonymous version of the dataset (considering all attributes as quasi-identifier), rather than on the original unprotected dataset. In particular, they showed that the information loss entailed by microaggregation is largely compensated by the reduction in the amount of noise to be added to achieve differential privacy.

4 Advanced Application Scenarios

While k-anonymity has been first formulated to protect the identities (and, as illustrated in the previous sections, further extended to also protect sensitive information) of the respondents involved in a microdata release, it has been adopted since then also in other application scenarios, possibly in extended/modified formulations as best suited for the considered context. In particular, in this section we discuss scenarios concerning location-based services and movements or positions of users (Section 4.1), social network analysis (Section 4.2), contact tracing (Section 4.3), and big data analytics (Section 4.4)

4.1 Location-Based Services and Movements Data

k-Anonymity has been proved to be effective in scenarios characterized by the private usage of location-based services, and by the private release of datasets of users' movements. We jointly discuss these scenarios as they are strictly interconnected, being both based on the release of (live or historical) spatial information of a set of individuals.

Location-Based Services. Location-Based Services (LBSs) enable users to obtain real-time services based on their current position: for example, an LBS can be used to query for nearby points of interests, such as the nearest open restaurant. The release of the position of a user, necessary for obtaining an LBS, and its link to the user's identity can however be considered sensitive and need to be protected. Effectively protecting users' privacy in the context of LBSs is a critical issue with multiple facets. On the one hand, location information of a user can be seen as quasi-identifying and can then be linked to external data sources to (re-)identify users. On the other hand, location information itself can be considered sensitive, and appropriate protection could be needed to hide the actual, precise location of a user. A third relevant facet relates to the actual content of a query submitted to an LBS, which may be sensitive regardless of if and how the identity or the location information of the submitting users are to be protected.

k-Anonymity can be effectively adopted to facilitate anonymous usage of LBSs. Among the first attempts to adopt k-anonymity in this context, Gruteser and Grunwald [38] put forward the idea of considering an LBS request sent by a user k-anonymous if it is indistinguishable from the spatial (and temporal) information of at least $k-1$ other requests sent from different users (so to associate any location-based request with at least k individuals). This intuition has been used by different solutions permitting anonymous usage of location-based services (e.g., [10, 36, 38, 55, 56]). The privacy requirement is typically enforced by decreasing the accuracy of the location information to be sent to the location-based service provider, blurring it to a *cloaked area* that includes at least k different individuals. In this way, the service provider (and any subject observing an LBS request) can match any request, through its location information, to at least k different users. It is intuitive that the rationale behind spatial cloaking shows similarities with the rationale behind generalization in traditional k-anonymity: the coarser the level of details of quasi-identifying information (be them attributes in database scenarios, or locations in LBSs), the harder the possibility to relate it to a single individual. Spatial cloaking is typically enforced by a *trusted entity* that acts as anonymizer, and mediates the communication between a user, requesting a location-based service, and the LBS provider. Instead of communicating directly to the provider, the user sends her request to the anonymizer, which modifies the (precise) location information included in the original request to obtain a cloaked region with at least k users, and forwards the request with the cloaked region to the provider. Clearly, being the provider's response based on a larger area, it can include spurious results that would not have been returned evaluating the actual, precise location of the user. Such results then can be directly filtered by the user. It is interesting to note that, to generate useful responses from the provider, cloaked areas must be *inclusive* (i.e., defined so to guarantee that all results that would be obtained on the precise real location of the user are also included among those returned for the cloaked area), and *minimal* (i.e., defined not to include more data items than necessary to ensure k-anonymity) [56]. Elaborating on these key concepts, different solutions have addressed specific aspects of different application scenarios, such as the private evaluation of location-based queries on cloaked regions [55] and on constrained spaces such as road networks (which inevitably complicate spatial cloaking as users' po-

sitions are often constrained by roads and intersections) [56], the support for personalized privacy preferences (e.g., the support for different values of k specified by users) [36] possibly in specific scenarios such as autonomous vehicles in cyber-physical systems [75], the support for distributed location anonymizers to limit the risk of being a single point of failure [10], the protection of location information in crowdsensing scenarios [78], and the protection of location information when users request navigation routes [48].

A slight modification of k-anonymity, which however shares with it the main idea of protection through hiding in a crowd, is the concept of *feeling-based* location privacy [83], according to which a location of a user can be safely released to an LBS provider if it is at least popular (with the popularity of a place measured in terms of the entropy of the footprints left by users in the place) as that of a public place specified by the user as baseline. Spatial cloaking is then used to blur the actual location of each user into a (larger) area that satisfies this privacy requirement.

Movement data publication. Another application scenario in the context of location data that has seen the adoption of k-anonymity concerns the privacy-preserving publication of movement datasets. Publishing or sharing movements, or trajectories, followed by users may prove useful to derive knowledge about, for example, how individuals move within a region, or whether road networks or public transportation systems should be improved. To protect the privacy of the individuals involved in the release, different privacy requirements and notions for trajectories based on k-anonymity have then been proposed. One of the first attempts in this regard is the notion of (k, δ)-anonymity [1] where δ is the radius of a circular area surrounding each point of a trajectory (due to errors and imprecisions of the location measurements). (k, δ)-Anonymity is satisfied if there are at least k (anonymized) trajectories that follow the same route with respect to the value of δ (i.e., such that the i^{th} points of all trajectories fall within a circular area based on the value of δ). Given a value for δ, a (k, δ)-anonymous trajectory dataset can then be achieved by possibly translating some points of the original trajectories to be published [1]. Another notion of *trajectory* k-anonymity is proposed in [31]. This solution aims at counteracting the risk that a recipient can leverage her knowledge of some locations visited by a user to identify her trajectory in a de-identified trajectory data collection, and hence discover other locations visited by that user. The trajectory k-anonymity requirement demands that recipients can map an original trajectory they partially know to no less than k different anonymized trajectories. Trajectories are therefore grouped, based on their similarity, in clusters of at least k elements each, and are transformed perturbing their locations. Orthogonally to these privacy notions, the approach in [74] focuses on the problem of effectively generalizing location data represented through GPS coordinates.

4.2 Social Network Analysis

Publishing social network data can undoubtedly represent a useful way to gain insights on the social relations existing within a community of interest. Social network data are usually released in the form of graph structures where vertices represent the social network users, and edges represent existing relationships among them. The negative impacts that an uncontrolled release of such data can have on the privacy of the users in the social network are easy to understand, and have fostered lines of research aiming at permitting the release of *sanitized* social network graphs where users' personal information and/or relations are protected. Figure 10(a) illustrates a sample social network graph with 7 users (vertices), where the identities of users have been removed from the vertices. Backstrom et at. [8]

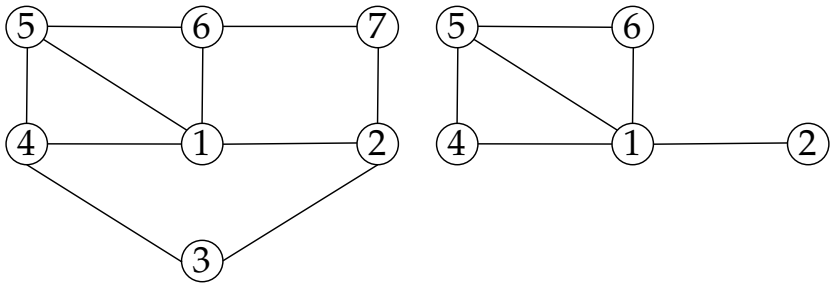


Figure 10: An example of a de-identified social network graph (a), and of the 1-neighborhood of de-identified vertex 1 (b)

observed that such a simple vertex de-identification does not provide any guarantee of anonymity. As pointed in [12], proper modifications to the graph structure (e.g., addition and deletion of vertices and/or edges to/from the graph, or grouping them into super-vertices and/or super-edges) are needed to ensure a desirable degree of protection. A first possibility is a random application of such modifications, which can provide some form of probabilistic resistance to re-identification [12]. Other solutions are based on k-anonymity to offer quantifiable protection against re-identification. A common assumption in this context considers the degree of the vertices of a social network graph as quasi-identifying information [39]. Building on this assumption, Liu and Terzi [51] propose the notion of *k-degree anonymity*, whose privacy requirement demands that each vertex in a graph has the same degree of at least other $k-1$ vertices in the graph. Given a social network graph, their proposal builds a sanitized version satisfying k-degree anonymity for a given value of k by adding a (minimal) set of edges to the input graph without modifying vertices. The notion of k-degree anonymity has then been adopted in other proposals addressing specific aspects of the problem, such as the protection of specific individuals in a social network (e.g., represented by structurally important vertices in the graph) [57]. k-Degree anonymity is also adopted by Casas-Roma et at. [11] who propose an approach for generating k-degree anonymous social network graphs combining edge removal and edge addition (possibly adopted together, resulting in a switching between edges), and considering a notion of edge relevance (defined in terms of neighborhood centrality of the graph edges) to improve the utility of the sanitized graph. Other proposals have investigated the possibility of computing k-degree anonymous social network graphs through modifications to the vertices of the original graph, such as the insertion of dummy vertices to represent dummy users of the social network (e.g., [15, 52]).

Zhou and Pei [87] propose a different privacy requirement, building on the assumption that a de-identified social network user may be re-identified through the structure of the 1-neighborhood sub-graph of the vertex representing the user. Figure 10(b) illustrates the 1-neighborhood of vertex 1 of the network in Figure 10(a). The rationale behind this assumption is that the 1-neighborhood sub-graph of the vertex representing a user actually represents information about her social relationships that recipients may know, such as the fact that she has only two friends that know each other. They propose the notion of *k-neighborhood anonymity*, whose privacy requirement requires the 1-neighborhood sub-graph of each vertex in the graph to be isomorphic to that of at least $k-1$ other vertices. They propose a method to enforce such privacy requirement based on the addition of a

set of edges in the social network graph. In case original vertices are labeled with properties (e.g., attributes of the represented entities such as the date of birth of users), edge addition can be coupled by the generalization of such properties (e.g., generalizing the actual birth dates by only releasing the years of birth) [87]. This approach has been extended in [71] to consider cases in which the recipient has knowledge of the social relationships of users that go beyond the first level (i.e., of vertices in the higher hops of a vertex). Variations/enhancements of these approaches have been investigated to address specific adversarial scenarios and re-identification risks (e.g., [7, 40, 42, 54, 85, 72, 88]).

4.3 Contact Tracing

Another application scenario that has seen the adoption of k-anonymity is contact tracing, which has gained momentum at the time of writing due to the recent COVID-19 pandemic (e.g., [59]), but whose scientific and societal values are not limited to this pandemic. In fact, it is well-known that close contacts among individuals are critical to the spread of any airborne disease. Contact tracing is then one possibility out of a toolbox of approaches for limiting the diffusion of an airborne disease, as it permits an early detection of possible contagion events. Typically, contact tracing is done through ad-hoc mobile apps, which send (e.g., broadcast) beacons that are intercepted by nearby devices. In this way, users' devices maintain a log of close users. Typical approaches have leveraged the Bluetooth Low Energy (BLE) interface of smartphones. A well-known contact tracing protocol developed by Google and Apple protects the privacy of a user by encrypting her beacon IDs with temporal daily keys that can be disseminated in the event she tested positive [73]. Ali and Dyo [5] propose an approach based on the hashing of the representations of contacts among users, while achieving k-anonymity. Tedeschi et al. [68] propose a different contact tracing architecture, based on IoT, where users' smartphones only send (and never collect) BLE beacons to IoT *totems*, smart devices equipped with BLE transceivers. Totems collect beacons from users' devices and forward them to a central authority. When a user, recorded at a certain time at totem t, tests positive, the central authority publishes *all* tokens (i.e., regardless of whether the associated users tested positive or negative) received from t close at that time. Each user of the app then locally checks whether her beacon is in the list. As pointed out by the authors of [68], this solution achieves k-anonymity with respect to disclosing diagnosis: given a dataset of k beacons from totem t published by the central authority, each beacon could be the positive one (with a probability of $1/k$ of being the actual positive one). The adoption of geolocation data instead of BLE beacons has also been investigated for contact tracing: proposals in this direction (e.g. [13, 37, 41]) can protect respondents' privacy through more traditional k-anonymity-based solutions developed, for example, in the context of location data.

In the context of studying or predicting the evolution of airborne diseases, also offline analysis of human mobility data can be crucial for preventing or containing outbreaks [4]. Protecting the privacy of the involved individuals is clearly a key requirement in these scenarios [43]. For example, Chang et al. [14] studied the hourly movements of 98 million users from neighborhoods to points of interest (such as restaurants and religious establishments) to simulate the spread of COVID-19 in large metropolitan areas. In their analysis, they use location data from mobile applications and, to provide a basic protection to the individuals involved in the analysis, locations where fewer than five mobile devices were recorded are excluded (thus following a naive application of the 5-anonymity privacy requirement).

4.4 Big Data Analytics

We close this section with some observations on the adoption of k-anonymity for permitting private analytics of big data. While there is not a unique definition of big data, five distinctive characteristics that are usually adopted for defining big data are *volume*, *variety*, *velocity*, *veracity*, and *value*. Volume refers to the fact that the size of big data collections is typically very large. Variety refers to the fact that big data collections usually contain data collected from different sources. Velocity refers to the fact that big data are usually generated, collected, and analyzed at a fast pace, possibly in streams. Veracity refers to the quality and accuracy of big data. Value refers to the value that big data can provide. In principle, as observed in [84], the *volume* of big data may represent an advantage in the context of privacy (and, most specifically, of privacy through k-anonymity): intuitively, the more the records in a dataset, the more the respondents that can be used for 'hiding' the individual identities and, for example in the context of generalization-based approaches, the lesser the amount of generalization that can be necessary to hide identities in groups. However, effectively anonymizing big data entails issues that need careful analysis. Soria-Comas and Domingo-Ferrer [65] identify three properties of privacy models, including those based on k-anonymity, that should be evaluated for assessing their applicability to big data: *composability* (required for preserving privacy), *computational cost*, and *linkability* (demanded for permitting analytics over multiple data sources). In a nutshell, composability concerns the guarantee that independent repeated applications of a privacy model preserve the privacy guarantees ensured by the model, so that multiple and independent releases of private data satisfy the privacy requirement even when considered in combination [35]. The computational cost of a privacy model estimates how much work is needed to transform the original data collection in a protected one that satisfies the privacy requirement of the adopted model. Linkability concerns the possibility of linking multiple records related to the same individual (possibly in different data collections to be anonymized) [65].

- As for the *composability* property, k-anonymity has been specifically designed to anonymize a single dataset. If applied to datasets with overlapping respondents, k-anonymity may not guarantee composability [67]. However, as noted in [65], k-anonymity can preserve composability if no overlapping respondents nor sensitive attributes are included in the different (anonymized) datasets or when, if different sources include the same respondents, equivalence classes across the datasets include the same respondents [27].
- As for the *computation cost*, the problem of computing an optimal k-anonymous dataset (e.g., minimizing generalization or optimizing microaggregation) is computationally hard [62, 65]. However, it has to be noted that there exist (heuristic) approaches that can be adopted to compute (approximate) k-anonymous solutions in reasonable time. For example, Incognito (Section 3.1) is experimentally proved to ensure good performances when QI attributes are limited in number [44]. Mondrian (Section 3.1) adopts a heuristic approach quasi-linear in the number n of records of the dataset to be anonymized ($O(n \log n)$) [45]. MDAV, an enforcement approach that adopts microaggregation (Section 3.2), has quadratic cost in the number of records in the dataset to be anonymized ($O(n^2)$) [30]. Hence, existing approaches may be applied to datasets composed of a large number of records as expected in big data scenarios.
- As for the *linkability* property, a k-anonymous relation always permits to determine the equivalence class to which a respondent (for which quasi-identifying values are

known) belongs. If a recipient is included in two or more k-anonymous relations, it is then possible to (at least) link the equivalence classes to which she and for this reason, to a certain degree, k-anonymity guarantees linkability. It is however to be noted that, although linkability is a desideratum in the context of big data, the more the degree of linkability, the more the amount of information about respondents that is leaked and, as mentioned above, the possibility for a recipient to belong to multiple k-anonymous relations makes k-anonymity not always composable.

The observations above suggest that a definitive answer to whether k-anonymity can be effectively adopted in the context of the anonymization of big data is still to be given. Considering its computational cost, leveraging k-anonymity for anonymizing big data can successfully deal with their *volume* and *velocity*. Specific large-scale and distributed computation paradigms, such as MapReduce or solutions based on Apache Spark, can also be adopted for ensuring scalability and further improve performances of the computation of k-anonymous solutions (e.g., [22, 23, 84]). On the other hand, its limited composability (and, similarly, its linkability property) may represent an issue with the *variety* of big data when data from multiple k-anonymous sources are combined, with the risk that such combination satisfies k'-anonymity with $k'<k$ (and, in the worst case scenario, $k'=1$). There are preliminary attempts for adopting, and adapting, k-anonymity to big data anonymization. Salas and Torra [61] propose a microaggregation-based approach for computing k-anonymity over evolving relations, published over time, that ensures composability. In particular, composability is guaranteed by the fact that all relations to be published over time are managed by a single data owner, who therefore has full visibility over the data and over their evolutions over time [61].

Domingo-Ferrer and Soria-Comas [28] observe that, in the context of big data, the traditional distinction between quasi-identifying attributes and confidential attributes may blur, especially when different datasets come from different (and possibly not fully trusted) data owners. In this case, they observe that an untrusted owner may share, leak or sell any confidential attribute of the respondents of its dataset, so that these confidential attributes may become part of the (identified) information available for mounting re-identification attacks – or, similarly, use it herself to re-identify the related respondents coming from other k-anonymous datasets. This observation is in line with the fact that in big data scenarios, information can be generated and released by a multitude of subjects (in contrast to a few data collectors/publishers such as statistical agencies that characterized more traditional scenarios). Simply adding all sensitive attributes to the quasi-identifier and adopt k-anonymity would unfortunately –especially when the number of sensitive attributes is high– considerably increase the size of the quasi-identifier hence resulting, as pointed in [2], in a large amount of generalization and consequent large amount of data loss. The approach in [28], focusing on the anonymization of a single k-anonymous relation where sensitive attributes can be quasi-identifying, tackles this issue by computing and releasing multiple k-anonymous views over the original relation. In particular, given a relation $R(a^q_1, \ldots, a^q_n, a^c_1, \ldots, a^c_m)$ to be anonymized, where a^q_i is the i^{th} quasi-identifying attribute and a^c_i is the i^{th} confidential attribute that could also be used as quasi-identifying, the approach in [28] proposes to release a set of m k-anonymous relations, where the i^{th} relation ($i = 1, \ldots, m$) has schema $R(a^q_1, \ldots, a^q_n, a^c_i)$. If confidential attributes are non-correlated, then these m k-anonymous relations can be independently computed adopting any k-anonymity algorithm considering a^c_i among the quasi-identifier, $i=1, \ldots, m$. If, on the other hand, confidential attributes are correlated, they may be linked (with a certain degree of precision) among the different k-anonymous datasets, permitting intersection attacks and

possibly pinpointing specific respondents as the independent generalizations of the different datasets may cluster records (and hence respondents) differently. When this is the case, the generalization can be performed in two steps: 1) attributes $a_1^q, \ldots, a_n^q$ are first generalized to obtain a set of k-anonymous equivalence classes, 2) the i^{th} relation corresponds to the set of equivalence classes computed in the first step and the generalization of the values of attribute a_i^c within each class. This permits to create, among the different k-anonymous views, the same equivalence classes (i.e., an equivalence class contains the same generalized records in all views, as in all views they are computed generalizing $a_1^q, \ldots, a_n^q$).

5 Conclusions

In this paper, we have discussed the theory and application of k-anonymity. We have illustrated the main privacy requirements pursued by k-anonymity and by some of its well-known extensions, and we have presented enforcement approaches and algorithms based on data generalization, data fragmentation, and microaggregation. As highlighted in the paper, k-anonymity still represents a valid approach for addressing some aspects of the complex privacy problem, and several and diverse application scenarios have recently seen (and more will reasonably see) the adoption of k-anonymity. We have also discussed some of these application scenarios, including location-based services, movement data publication, analysis of social network data, contact tracing applications, and big data analytics.

Acknowledgments

This work was supported in part by the EC within the H2020 Program under projects MOSAICrOWN and MARSAL, by the Italian Ministry of Research within the PRIN program under project HOPE, and by JPMorgan Chase & Co under project "k-anonymity for AR/VR and IoT/5G".

References

[1] O. Abul, F. Bonchi, and M. Nanni. Never walk alone: Uncertainty for anonymity in moving objects databases. In *Proc. of ICDE 2008*, pages 376–385, Cancun, Mexico, April 2008.

[2] C. C. Aggarwal. On k-anonymity and the curse of dimensionality. In *Proc. of VLDB 2005*, pages 901–909, Trondheim, Norway, August/September 2005.

[3] G. Aggarwal, M. Bawa, P. Ganesan, H. Garcia-Molina, K. Kenthapadi, R. Motwani, U. Srivastava, D. Thomas, and Y. Xu. Two can keep a secret: A distributed architecture for secure database services. In *Proc. of CIDR 2005*, pages 186–200, Asilomar, CA, USA, January 2005.

[4] L. Alessandretti. What human mobility data tell us about COVID-19 spread. *Nature Reviews Physics*, 4(1):12–13, 2022.

[5] J. Ali and V. Dyo. Cross hashing: Anonymizing encounters in decentralised contact tracing protocols. In *Proc. of ICOIN 2021*, pages 181–185, Jeju Island, Korea, January 2021.

[6] H.S. Asif, P.A. Papakonstantinou, and J. Vaidya. How to accurately and privately identify anomalies. In *Proc. of ACM CCS 2019*, pages 719–736, London, UK, November 2019.

[7] R. Assam, M. Hassani, M. Brysch, and T. Seidl. (k, δ)-Core anonymity: Structural anonymization of massive networks. In *Proc. of SSDBM 2014*, pages 1–12, Aalborg, Denmark, June/July 2014.

[8] L. Backstrom, C. Dwork, and J. Kleinberg. Wherefore art thou R3579X? anonymized social networks, hidden patterns, and structural steganography. In *Proc. of WWW 2007*, pages 181–190, Banff, Canada, May 2007.

[9] R. J. Bayardo and R. Agrawal. Data privacy through optimal k-anonymization. In *Proc. of ICDE 2005)*, pages 217–228, Tokyo, Japan, April 2005.

[10] F. Buccafurri, V. De Angelis, M. F. Idone, and C. Labrini. A distributed location trusted service achieving k-anonymity against the global adversary. In *Proc. of MDM 2021*, pages 133–138, Paphos, Cyprus, June 2021.

[11] J. Casas-Roma, J. Herrera-Joancomartí, and V. Torra. k-Degree anonymity and edge selection: Improving data utility in large networks. *KAIS*, 50(2):447–474, 2017.

[12] J. Casas-Roma, J. Herrera-Joancomartí, and V. Torra. A survey of graph-modification techniques for privacy-preserving on networks. *Artificial Intelligence Review*, 47(3):341–366, 2017.

[13] C. A. Cassa, S. J. Grannis, J. M. Overhage, and K. D. Mandl. A context-sensitive approach to anonymizing spatial surveillance data: impact on outbreak detection. *JAMIA*, 13(2):160–165, 2006.

[14] S. Chang, E. Pierson, P. W. Koh, J. Gerardin, B. Redbird, D. Grusky, and J. Leskovec. Mobility network models of COVID-19 explain inequities and inform reopening. *Nature*, 589(7840):82–87, 2021.

[15] 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. *SNAM*, 3(3):381–399, 2013.

[16] V. Ciriani, S. De Capitani di Vimercati, S. Foresti, and P. Samarati. k-Anonymity. In T. Yu and S. Jajodia, editors, *Secure Data Management in Decentralized Systems*, pages 323–353. Springer-Verlag, 2007.

[17] V. Ciriani, S. De Capitani di Vimercati, S. Foresti, and P. Samarati. Theory of privacy and anonymity. In M. Atallah and M. Blanton, editors, *Algorithms and Theory of Computation Handbook (2nd edition)*. CRC Press, 2009.

[18] C. Clifton and T. Tassa. On syntactic anonymity and differential privacy. In *Proc. of ICDEW 2013*, pages 88–93, Brisbane, Australia, April 2013.

[19] G. Cormode, C. M. Procopiuc, E. Shen, D. Srivastava, and T. Yu Empirical privacy and empirical utility of anonymized data. In *Proc. of ICDEW 2013*, pages 77–82, Brisbane, Australia, April 2013.

[20] C. Cumby and R. Ghani. A machine learning based system for semi-automatically redacting documents. In *Proc.of AAAI 2011*, pages 1628–1635, San Francisco, CA, USA, August 2011.

[21] S. De Capitani di Vimercati, R.F. Erbacher, S. Foresti, S. Jajodia, G. Livraga, and P. Samarati. Encryption and fragmentation for data confidentiality in the cloud. In A. Aldini, J. Lopez, and F. Martinelli, editors, *Foundations of Security Analysis and Design VII*, pages 212–243. Springer, 2014.

[22] S. De Capitani di Vimercati, D. Facchinetti, S. Foresti, G. Oldani, S. Paraboschi, M. Rossi, and P. Samarati. Scalable distributed data anonymization. In *Proc. of PerCom 2021*, pages 401–403, Kassel, Germany, March 2021.

[23] S. De Capitani di Vimercati, D. Facchinetti, S. Foresti, G. Oldani, S. Paraboschi, M. Rossi, and P. Samarati. Artifact: Scalable distributed data anonymization. In *Proc. of PerCom 2021*, pages 401–403, Kassel, Germany, March 2021.

[24] S. De Capitani di Vimercati, S. Foresti, S. Jajodia, G. Livraga, S. Paraboschi, and P. Samarati. Fragmentation in presence of data dependencies. *IEEE TDSC*, 11(6):510–523, 2014.

[25] S. De Capitani di Vimercati, S. Foresti, S. Jajodia, G. Livraga, S. Paraboschi, and P. Samarati. Loose associations to increase utility in data publishing. *JCS*, 23(1):59–88, 2015.

[26] S. De Capitani di Vimercati, S. Foresti, G. Livraga, and P. Samarati. Data privacy: Definitions

and techniques. *IJUFKBS*, 20(6):793–817, 2012.

[27] J. Domingo-Ferrer. Big data anonymization requirements vs privacy models. In *Proc. of SECRYPT 2018*, pages 471–478, Porto, Portugal, July 2018.

[28] J. Domingo-Ferrer and J. Soria-Comas. Anonymization in the time of big data. In *Proc. of PSD 2016*, pages 57–68, Dubrovnik, Croatia, September 2016.

[29] J. Domingo-Ferrer, J. Soria-Comas, and R. Mulero-Vellido. Steered microaggregation as a unified primitive to anonymize data sets and data streams. *IEEE TIFS*, 14(12):3298–3311, 2019.

[30] J. Domingo-Ferrer and V. Torra. Ordinal, continuous and heterogeneous k-anonymity through microaggregation. *Data Mining and Knowledge Discovery*, 11(2):195–212, 2005.

[31] J. Domingo-Ferrer and R. Trujillo-Rasua. Microaggregation-and permutation-based anonymization of movement data. *Information Sciences*, 208:55–80, 2012.

[32] C. Dwork. Differential privacy. In *Proc. of ICALP 2006*, pages 1–12, Venice, Italy, July 2006.

[33] Federal Committee on Statistical Methodology. *Statistical policy working paper 22 (Second Version)*. USA, December 2005. Report on Statistical Disclosure Limitation Methodology.

[34] K. B. Frikken and Y. Zhang. Yet another privacy metric for publishing micro-data. In *Proc. of WPES 2008*, pages 117–122, Alexandria, VA, USA, October 2008.

[35] S. R. Ganta, S. P. Kasiviswanathan, and A. Smith. Composition attacks and auxiliary information in data privacy. In *Proc. of KDD 2008*, pages 265–273, Las Vegas, NV, USA, August 2008.

[36] B. Gedik and L. Liu. Protecting location privacy with personalized k-anonymity: Architecture and algorithms. *IEEE TMC*, 7(1):1–18, 2008.

[37] J. González-Cabañas, Á. Cuevas, R. Cuevas, and M. Maier. Digital contact tracing: Large-scale geolocation data as an alternative to bluetooth-based apps failure. *Electronics*, 10 (article 1093), 2021.

[38] M. Gruteser and D. Grunwald. Anonymous usage of location-based services through spatial and temporal cloaking. In *Proc. of MobiSys 2003*, pages 31–42, San Francisco, CA, USA, May 2003.

[39] M. Hay, G. Miklau, D. Jensen, D. Towsley, and P. Weis. Resisting structural re-identification in anonymized social networks. In *Proc. of VLDB 2008*, pages 102–114, Auckland, New Zealand, August 2008.

[40] X. He, J. Vaidya, B. Shafiq, N. Adam, and V. Atluri. Preserving privacy in social networks: A structure-aware approach. In *Proc. of WI-IAT 2009*, pages 647–654, Milan, Italy, September 2009.

[41] R. Iyer, R. Rex, K. P. McPherson, D. Gandhi, A. Mahindra, A. Singh, and R. Raskar. Spatial k-anonymity: A privacy-preserving method for COVID-19 related geospatial technologies. In *Proc. of GISTAM 2021*, pages 75–81, virtual, April 2021.

[42] H. Jiang, J. Yu, X. Cheng, C. Zhang, B. Gong, and H. Yu. Structure-attribute-based social network deanonymization with spectral graph partitioning. *IEEE TCSS*, pages 1–12, 2021. To appear.

[43] G. Jung, H. Lee, A. Kim, and U. Lee. Too much information: Assessing privacy risks of contact trace data disclosure on people with COVID-19 in South Korea. *Frontiers in Public Health*, 8 (article 305), 2020.

[44] K. LeFevre, D.J. DeWitt, and R. Ramakrishnan. Incognito: Efficient full-domain k-anonymity. In *Proc. of SIGMOD 2005*, pages 49–60, Baltimore, MD, USA, June 2005.

[45] K. LeFevre, D.J. DeWitt, and R. Ramakrishnan. Mondrian multidimensional k-anonymity. In *Proc. of ICDE 2006*, Atlanta, GA, USA, April 2006.

[46] F. Li, J. Sun, S. Papadimitriou, G.A. Mihaila, and I. Stanoi. Hiding in the crowd: Privacy preservation on evolving streams through correlation tracking. In *Proc. of ICDE 2007*, pages 686–695, Istanbul, Turkey, April 2007.

[47] L. Li, B. Pal, J. Ali, N. Sullivan, R. Chatterjee, and T. Ristenpart. Protocols for checking compro-

mised credentials. In *Proc. of CCS 2019*, pages 1387–1403, London, UK, November 2019.

[48] M. Li, Y. Chen, N. Kumar, C. Lal, M. Conti, and M. Alazab. Quantifying location privacy for navigation services in sustainable vehicular networks. *IEEE TGCN*, 2022. To appear.

[49] N. Li, T. Li, and S. Venkatasubramanian. t-closeness: Privacy beyond k-anonymity and ℓ-diversity. In *Proc. of ICDE 2007*, pages 106–115, Istanbul, Turkey, 2007.

[50] N. Li, W. Qardaji, and D. Su. On sampling, anonymization, and differential privacy or, k-anonymization meets differential privacy. In *Proc. of ASIACCS 2012*, pages 32–33, Seoul, Korea, May 2012.

[51] K. Liu and E. Terzi. Towards identity anonymization on graphs. In *Proc. of SIGMOD 2008*, pages 93–106, Vancouver, Canada, June 2008.

[52] T. Ma, Y. Zhang, J. Cao, J. Shen, M. Tang, Y. Tian, A. Al-Dhelaan, and M. Al-Rodhaan. KDVEM: A k-degree anonymity with vertex and edge modification algorithm. *Computing*, 97(12):1165–1184, 2015.

[53] A. Machanavajjhala, D. Kifer, J. Gehrke, and M. Venkitasubramaniam. ℓ-diversity: Privacy beyond k-anonymity. *ACM TKDD*, 1(1):3:1–3:52, March 2007.

[54] S. Mauw, Y. Ramírez-Cruz, and R. Trujillo-Rasua. Conditional adjacency anonymity in social graphs under active attacks. *KAIS*, 61(1):485–511, 2019.

[55] M. F. Mokbel, C.-Y. Chow, and W. G Aref. The new Casper: Query processing for location services without compromising privacy. In *Proc. of VLDB 2006*, pages 763–774, Seoul, Korea, September 2006.

[56] K. Mouratidis and M. L. Yiu. Anonymous query processing in road networks. *IEEE TKDE*, 22(1):2–15, 2009.

[57] F. Nagle, L. Singh, and A. Gkoulalas-Divanis. EWNI: efficient anonymization of vulnerable individuals in social networks. In *Proc. of PAKDD 2012*, pages 359–370, Kuala Lumpur, Malaysia, May 2012.

[58] M.E. Nergiz, C. Clifton, and A.E. Nergiz. Multirelational k-anonymity. In *Proc. of ICDE 2007*, pages 1417–1421, Istanbul, Turkey, April 2007.

[59] J. Park, E. Ahmed, H.S. Asif, J. Vaidya, and V.K. Singh. Privacy attitudes and COVID symptom tracking apps: Understanding active boundary management by users. In *Proc. of iConference 2022*, pages 332–346, virtual, February-March 2022.

[60] J. Salas and J. Domingo-Ferrer. Some basics on privacy techniques, anonymization and their big data challenges. *Mathematics in Computer Science*, 12(3):263–274, 2018.

[61] J. Salas and V. Torra. A general algorithm for k-anonymity on dynamic databases. In *Proc. of DPM and CBT 2018*, pages 407–414, Barcelona, Spain, September 2018.

[62] P. Samarati. Protecting respondents' identities in microdata release. *IEEE TKDE*, 13(6):1010–1027, November/December 2001.

[63] E. Shmueli and T. Tassa. Privacy by diversity in sequential releases of databases. *Information Sciences*, 298:344–372, 2015.

[64] J. Soria-Comas and J. Domingo-Ferrer. Probabilistic k-anonymity through microaggregation and data swapping. In *Proc. of FUZZ-IEEE 2012*, pages 1–8, Brisbane, Australia, June 2012.

[65] J. Soria-Comas and J. Domingo-Ferrer. Big data privacy: challenges to privacy principles and models. *Data Science and Engineering*, 1(1):21–28, 2016.

[66] J. Soria-Comas, J. Domingo-Ferrer, D. Sánchez, and S. Martínez. Enhancing data utility in differential privacy via microaggregation-based k-anonymity. *The VLDB Journal*, 23(5):771–794, 2014.

[67] K. Stokes and V. Torra. Multiple releases of k-anonymous data sets and k-anonymous relational databases. *IJUFKBS*, 20(6):839–853, 2012.

[68] P. Tedeschi, S. Bakiras, and R. Di Pietro. IoTrace: A flexible, efficient, and privacy-preserving IoT-enabled architecture for contact tracing. *IEEE Communications Magazine*, 59(6):82–88, 2021.

[69] M. Terrovitis, N. Mamoulis, and P. Kalnis. Privacy-preserving anonymization of set-valued data. *PVLDB*, 1(1):115–125, August 2008.

[70] M. Thouvenot, O. Curé, and P. Calvez. Preventing attribute and entity disclosures: Combining k-anonymity and Anatomy over RDF graphs. In *Proc. of Big Data 2021*, pages 5460–5469, virtual, December 2021.

[71] B.K. Tripathy and G.K. Panda. A new approach to manage security against neighborhood attacks in social networks. In *Proc. of ASONAM 2010*, pages 264–269, Odense, Denmark, August 2010.

[72] N.M. Uplavikar, J. Vaidya, D. Lin, and W. Jiang. Privacy-preserving friend recommendation in an integrated social environment. In *Proc. of ICISS 2020*, pages 117–136, Jammu, India, December 2020.

[73] S. Vaudenay. Analysis of DP3T - Between Scylla and Charybdis. Cryptology ePrint Archive, Report 2020/399, 2020. `https://ia.cr/2020/399`.

[74] J. Verdonck, K. De Boeck, M. Willocx, J. Lapon, and V. Naessens. A clustering approach to anonymize locations during dataset de-identification. In *Proc. of ARES 2021*, pages 1–10, virtual, August 2021.

[75] J. Wang, Z. Cai, and J. Yu. Achieving personalized k-anonymity-based content privacy for autonomous vehicles in CPS. *IEEE TII*, 16(6):4242–4251, 2019.

[76] K. Wang and B.C.M. Fung. Anonymizing sequential releases. In *Proc. of KDD 2006*, pages 414–423, Philadelphia, PA, USA, August 2006.

[77] K. Wang, Y. Xu, R. Wong, and A. Fu. Anonymizing temporal data. In *Proc. of ICDM 2010*, pages 1109–1114, Sydney, Australia, December 2010.

[78] X. Wang, Z. Liu, X. Tian, X. Gan, Y. Guan, and X. Wang. Incentivizing crowdsensing with location-privacy preserving. *IEEE TWC*, 16(10):6940–6952, 2017.

[79] R. C.-W. Wong, J. Li, A. Fu, and K. Wang. (α, k)-Anonymity: An enhanced k-anonymity model for privacy preserving data publishing. In *Proc. of KDD 2006*, pages 754–759, Philadelphia, PA, USA, August 2006.

[80] X. Xiao and Y. Tao. Anatomy: Simple and effective privacy preservation. In *Proc. of VLDB 2006*, pages 139–150, Seoul, Korea, September 2006.

[81] X. Xiao and Y. Tao. Personalized privacy preservation. In *Proc. of SIGMOD 2006*, pages 229–240, Chicago, IL, USA, June 2006.

[82] X. Xiao and Y. Tao. m-invariance: Towards privacy preserving re-publication of dynamic datasets. In *Proc. of SIGMOD 2007*, pages 689–700, Beijing, China, June 2007.

[83] T. Xu and Y. Cai. Feeling-based location privacy protection for location-based services. In *Proc. of CCS 2009*, pages 348–357, Chicago, IL, USA, November 2009.

[84] H. Zakerzadeh, C. C. Aggarwal, and K. Barker. Privacy-preserving big data publishing. In *Proc. of SSBDM 2015*, pages 1–15, La Jolla, CA, USA, July 2015.

[85] X. Zheng, G. Luo, and Z. Cai. A fair mechanism for private data publication in online social networks. *IEEE TNSE*, 7(2):880–891, 2020.

[86] B. Zhou, Y. Han, J. Pei, B. Jiang, Y. Tao, and Y. Jia. Continuous privacy preserving publishing of data streams. In *Proc. of the EDBT 2009*, pages 648–659, Saint Petersburg, Russia, March 2009.

[87] B. Zhou and J. Pei. The k-anonymity and ℓ-diversity approaches for privacy preservation in social networks against neighborhood attacks. *KAIS*, 28(1):47–77, 2011.

[88] L. Zou, L. Chen, and M. T. Özsu. K-automorphism: A general framework for privacy preserving network publication. *PVLDB*, 2(1):946–957, 2009.