From k-anonymity to Differential Privacy: A Brief Introduction to Formal Privacy Models

Muhammad Imran Khan, Simon N Foley, Barry O'Sullivan

From *k*-anonymity to Differential Privacy: A Brief Introduction to Formal Privacy Models

Muhammad Imran Khan[†], Simon N. Foley[‡], and Barry O'Sullivan[†]

[†]Insight Centre for Data Analytics, School of Computer Science and Information Technology, University College Cork, Ireland.
[‡]Department of Information Security and Communication Technology, Norwegian University of Science and Technology, Gjøvik, Norway.

Abstract

A number of formal privacy definitions also known as privacy models are presented when these definitions are followed then the anonymized data manifests some formal guarantees. There are several privacy definitions proposed in the literature including *k*-anonymity, differential privacy *l*-diversity, *t*-closeness so on and so forth. In this paper, we review some of the well-know formal privacy definitions.

Keywords— *k*-anonymity, *l*-diversity, *t*-closeness, (α, *k*)-anonymity, Multi-relational *k*-anonymity, ϵ-differential Privacy, Relational Database Management Systems, Privacy Definitions, Privacy Models.

1 Introduction: Countless Shades of Privacy

Privacy is a challenging concept to define because of its culturally subjective basis. A classic paper [1] considers the philosophical dimensions of privacy in three perspectives: *"i. a claim, entitlement or right of an individual to determine what information about himself may be communicated to others [2, 3]. ii. measure of the control an individual has over: (a) information about himself; (b) intimacies of personal identity; or (c) who has sensory access to himself [4, 5, 6, 7]. iii. state or condition of limited access to a person [8, 9]."* This right to privacy is a globally recognised right as stated in Article 12 of the UN's Universal Declaration of Human Rights that states [10],

> "*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.*"

Taking the privacy concerns of the modern-day digital world, legislators have provided legal cover to the public to ensure their right to privacy. Dominant examples of such legislation include the Health Insurance Portability and Accountability Act (HIPAA, 1996) [11], the EU's General Data Protection Regulation [12], the OECD privacy framework [13]. 58% of total countries in the world have data protection and privacy legislation in place while 10% have draft legislation and 21% of the countries still have no legislation in place (there is no data for the remaining 12%) [14].

Contemporary organizations collect large volumes of data over which analytics are routinely carried out for various purposes, including data-driven marketing and informed decision making. Organizations often delegate this task to third parties specializing in data analytics. The collection, storage, processing, and sharing of personal data raises dimensions privacy concerns. The work of Tore Dalenius in [15], points out the problem of publishing the data of populations without risking individual's privacy. The research in this domain started to gain momentum in the late 1970s and 1980s, especially in the context of publishing census data. Recently privacy research is no longer limited to census data, as emerging new technologies have brought privacy concerns along, for example, the privacy concerns in social networks [16, 17, 18], internet of things [19, 20, 21] and e-commerce [22, 23, 24].

The relational data model is the most widely used data model for storage and processing of data [25, 26]; for this reason, the relational data model in DBMS (Relational DBMS – RDBMS) is considered in this work. To ensure the privacy of individuals, the data is anonymized. The anonymized version of data is typically made available in two settings within the RDBMS framework, that are, non-interactive and interactive query setting. Interactive query setting allows dynamic queries, while in a non-interactive setting an anonymized version of the entire database is made available. In general, data is considered anonymized if it complies with a formal privacy definition. Several of the formal definitions of privacy have been proposed in the literature that are reviewed in the next Section 2. When an anonymized version of data is made available, then in that context an adversary aims to strive for identifying individuals and disclosing their sensitive attributes in the database.

The next section looks further into the formal definitions of privacy research proposed in the literature to achieve anonymization in the context of databases.

Section 2 examines well-known formal privacy definitions proposed in the literature to achieve anonymization in the context of databases. Section 3 presents the key observations gathered from the literature. Conclusions are drawn in Section 4.

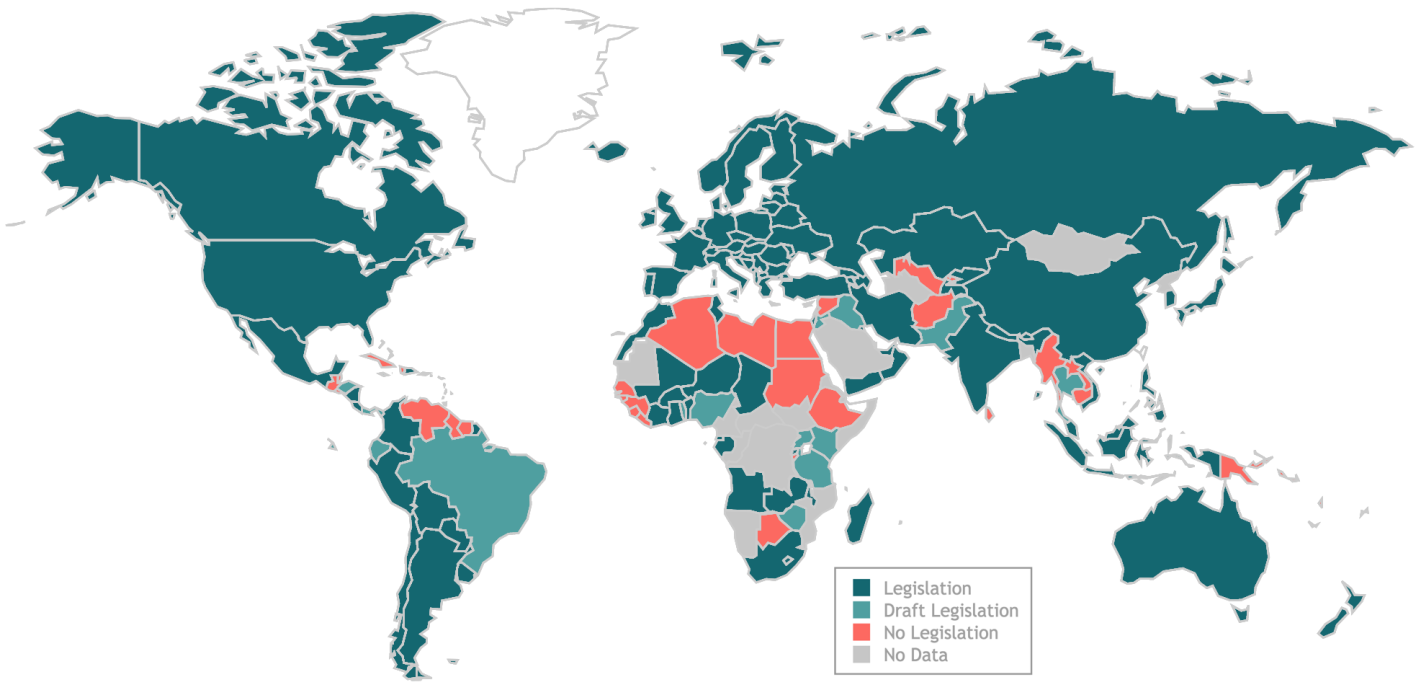

Figure 1: The state of Data protection and privacy legislation worldwide [14].

2 Formal Privacy Definitions

Several forms of privacy have been formalized in the literature. The two mainstream definitions of privacy are k-anonymity [27] and differential privacy [28]. k-anonymity serves as the foundation for several privacy definitions including l-diversity [29], t-closeness [30], (α, k)-anonymity [31]. This section reviews the well-known formal privacy definitions. The definitions are described within the framework of relational databases, and therefore, before reviewing the privacy definitions, the next section defines the key relational database terms used in this paper.

2.1 Key Relational Database Terms

In a relational model a relation instance is denoted as $\mathcal{T} = \{r_1, r_2, \ldots, r_n\}$ where r_i is a tuple of attribute values that represents a record. $\mathcal{T}$ is a subset of some larger population $\mathcal{U}$. Each tuple represents an individual from $\mathcal{U}$. Let the set of attributes be denoted by $Attr = \{attr_1, attr_2, \ldots, attr_n\}$, for example, the table shown in Figure 1, $Attr$ = {`Name`, `City`, `Country`}. The value of attribute $attr_j$ for a tuple r_i is denoted as $r_i[attr_j]$, for example, r_1[`Country`] = ('Spain'). $r_i[Attr]$ denotes the tuple $(r_i[attr_1], r_i[attr_2], \ldots, r_i[attr_n])$ which is the projection of r onto the attributes in $Attr$ for example r_1[`Name`, `City`, `Country`] = ('Nicolau', 'Barcelona', 'Spain').

r_i	`Name`	`City`	`Country`
r_1	Nicolau	Barcelona	Spain
r_2	Jordi	Barcelona	Spain
r_3	John	New York	USA
r_4	Sean	Cork	Ireland
r_5	Patrick	Dublin	Ireland
r_6	Matías	Berlin	Germany

Table 1: An example relation $\mathcal{T}$.

2.2 *k*-anonymity based and Extensions

k-anonymity [32] can be considered among the first formal definitions of privacy. The following sections presents k-anonymity and its extensions.

2.2.1 *k*-anonymity

In the context of k-anonymity, attributes are classified in the following non-exclusive categories, *Identifiers*, *Quasi-Identifiers*, and *Sensitive attributes*. The classification is typically performed based on the risk of record re-identification using these attributes and the sensitivity of the information these attributes convey.

- **Identifier:** an identifier is defined as *"an attribute that refers to only a particular individual in the given population $\mathcal{U}$"*. Let the set of identifiers be denoted as $\mathcal{I} \in Attr$. An example of an identifier is the Personal Public Service Number (PPS Number) which can uniquely identify individuals in Ireland. Other examples include an individual's passport number, driving license number, and e-mail address.
- **Quasi-Identifier (QI):** quasi-identifiers by themselves do not uniquely identify individuals; however, when correlated with other available external data, an individual (or individuals) can be identified. A quasi-identifier is defined in [32, 33] as a *"set of non-sensitive attributes of a relation if linked with external data to then uniquely identify at least one individual in the population $\mathcal{U}$"*. Let the set of quasi-identifiers be denoted as $QI \in Attr$. An example of quasi-identifier is the set of attributes `Zipcode`, `Date of Birth`, and `Gender`. For instance, the set of attributes `Zipcode`, `Date of Birth`, and `Gender` was used to re-identify governor of Massachusetts in [33]. The re-identification was performed by directly linking shared attributes in two datasets, i.e. voter rolls and insurance company datasets. It was reported that 87% of the US population could be identified by these three attributes [33].
- **Sensitive attribute:** sensitive attributes consist of sensitive person-specific information. This information includes salary, disability status, or disease. The set of sensitive attributes is denoted as $SenAttr \in Attr$. All possible values for the sensitive attribute is denoted as $SAV = \{sv_1, sv_2, \ldots, sv_n\}$.

k-anonymity is defined in [32] as follows, *"a relation $\mathcal{T}$ satisfies k-anonymity if and only if each tuple $r_i[QI] \in \mathcal{T}$ appears with at least k occurrence in $\mathcal{T}$"*.

k-anonymity [32] provides a degree of anonymity if the data for each person cannot be distinguished from k-1 individuals in a released dataset with respect to a set of quasi-identifiers. Given $QI \in Attr$ then two tuples r_i and r_j are quasi-identifier equivalent if $r_i[QI] = r_j[QI]$. The relation $\mathcal{T}$ can be divided into quasi-identifier equivalence classes. Let the set of all the equivalence classes in $\mathcal{T}$ be $\mathcal{E}$ where each equivalence class $e \in \mathcal{E}$ consists of all the rows that have the same values for each quasi-identifier.

Another way to define k-anonymity is that a relation $\mathcal{T}$ satisfies k-anonymity if the minimum equivalence class size is at least k in $\mathcal{T}$. Tables 2 and 3 show an original and a k-anonymized (3-anonymized) version of the microdata relations, respectively, where `Name` is an identifier, attributes `Age` & `Zipcode` are quasi-identifiers, and the attribute `Salary` is a sensitive attribute. In Table 3 the identifier `Name` is suppressed and is shown as * while quasi-identifiers `Age` & `Zipcode` are generalized – replacing with a semantically similar but less specific value.

In Table 3, records #1, #2, and #3, form an equivalence class, similarly records #4, #5, and #6 as well as records #7, #8, and #9 also forms equivalence classes with respect to the quasi-identifiers {`Age` & `Zipcode`}. Originally, k-anonymity was proposed for a one-time release of data, meaning that the user is not enabled to query the DBMS interactively.

Though considered to be among the first privacy definitions, k-anonymity, has been widely

applied in many domains to preserve privacy for examples Location-based services [34, 35, 36, 37, 38], ride-hailing services [39], and webmail auditing [40]. *k*-anonymity has been used along with cryptographic hashing to develop a protocol that provides a degree of anonymity while checking for passwords in a compromised databases [41].

#	Name	Age	Zipcode	Salary
1	Kenneth	23	3134	77k
2	Hendry	37	3135	77k
3	John	34	3134	83k
4	Noemi	52	7290	65k
5	James	58	7291	77k
6	Amanda	55	7290	83k
7	Miyu	49	3134	65k
8	Vlad	43	3135	65k
9	Alex	46	3134	83k

Table 2: A relation with `Name` as an identifier, `Age` & `Zipcode` are quasi-identifiers and `Salary` as a sensitive attribute.

#	Name	Age	Zipcode	Salary
1	*	<45	313*	77k
2	*	<45	313*	77k
3	*	<45	313*	83k
4	*	≥50	729*	65k
5	*	≥50	729*	77k
6	*	≥50	729*	83k
7	*	4*	313*	65k
8	*	4*	313*	65k
9	*	4*	313*	83k

Table 3: A 3-anonymized version of Table 2.

A weakness of *k*-anonymity is its susceptibility to the homogeneity attack and the background knowledge attack [29]. If all the values for one of the sensitive attributes, within an equivalence class, are same then it results in a homogeneity attack, for instance, if the adversary knows an individual who is 33 years old and lives in the zipcode 3134 and has record is in an equivalence class where all the salaries are 77k then the adversary deduces that the individual's salary is 77k, though the table is *k*-anonymized. In [29], it was shown that if the adversary knows that Japanese patients are less likely to have heart-related diseases (background knowledge) enabled an adversary to predict the diagonals for an individual.

2.2.2 *l*-diversity

Another well-known definition of privacy is *l*-diversity which is an extension of *k*-anonymity. The *l*-diversity principle in [29] states *"an equivalence class e is l-diverse if it contains at least*

'well-represented' values for the sensitive attribute $\in$ *SenAttr. A relation* $\mathcal{T}$ *is l-diverse if all the equivalence classes (q-block) are l-diverse"*. The term 'well-represented' is instantiated in three different ways as defined by the authors in [29] that are Distinct *l*-diversity, Entropy *l*-diversity, and Recursive *(c-l)*-diversity. The elementary form of *l*-diversity is distinct *l*-diversity. The remaining two instantiations are stronger instantiations and take the distribution of sensitive attribute values into account.

- Distinct *l*-diversity definition requires l distinct values in each equivalence class $e \in \mathcal{E}$.
- Entropy *l*-diversity is defined in [29] as *"A table l-diverse if for every equivalence class* $e \in \mathcal{E}$ the entropy of the distribution of sensitive values in each equivalence class is at least $log(l)$".
- Recursive (c, l)-diversity requires that the least frequent sensitive attribute values that do not appear rarely as well as most frequent sensitive attributes values do not appear too frequently in an equivalence class e.

Table 4 shows an *l*-diverse (3-diverse) version of Table 2, where each equivalence class has three distinct sensitive attribute values.

#	Name	age	Zipcode	Salary
7	*	<50	313*	65k
2	*	<50	313*	77k
3	*	<50	313*	83k
4	*	≥50	729*	65k
5	*	≥50	729*	77k
6	*	≥50	729*	83k
1	*	<50	313*	77k
8	*	<50	313*	65k
9	*	<50	313*	83k

Table 4: A *l*-diverse (3-diverse) version of Table 2.

In some scenarios it is difficult to achieve *l*-diversity. Let us say that a relation consisting of $20,000$ records has only one sensitive attribute, `Test Results`, where `Test Results` can take either negative or positive value. If 99.75% of the `Test Results` are positive, while 0.25% are negative, then there can be at most $20,000 * 0.25\% = 50$ equivalence classes.

l-diversity is susceptible to a *skewness attack* and *similarity attack* [30]. Consider a class with 9 positive values and 1 negative value, and another class with 9 negative values and 1 positive value, both classes are 2-diverse but present different privacy risks – skewness attack. *l*-diversity doesn't take the semantic-level closeness of the sensitive attribute values into account and thus leads to similarity attacks. For example, in a given equivalence class with the sensitive attribute `diagnosis`. Consider all the values for `diagnosis` in that given equivalence class are related to heart-related diseases. Therefore, an adversary knows someone in that given equivalence class then the adversary concludes that the individual has heart-related disease.

2.2.3 *t*-closeness

The *t*-closeness definition is a refinement of *l*-diversity. The authors in [30] define *t*-closeness as *"An equivalence class e is said to have t-closeness if the distance between the distribution of a sensitive attribute value in this class and the distribution of the sensitive attribute value in the whole relation $\mathcal{T}$ is no more than a threshold t."* As a distance metric, the Earth Mover's Distance can be used to measure the distance between two frequency distributions [30].

l-diversity treats sensitive attributes of all the equivalence classes in a similar manner without taking into account the global distribution of these sensitive attribute values in the relation. However, in the case of real-world datasets, sensitive attribute values might be skewed; therefore, it would be desirable to take into account the global distribution of sensitive attribute values. Limitations of *t*-closeness are discussed in [42], *t*-closeness degrades utility, is challenging to be achieved, and under certain scenarios, it is shown to be NP-hard [43].

2.2.4 (α, k)-anonymity

The (α, k)-anonymity is an enhanced version of k-anonymity. The (α, k)-anonymity defines that the frequency of sensitive attribute values α remains within the user-defined threshold in equivalence classes. The authors introduced an α-deassociation requirement in [31], if this requirement is satisfied along with k-anonymity, then it is said that (α, k)-anonymity is satisfied. Given a sensitive attribute value sv_i, a relation $\mathcal{T}$ is α-deassociated if the relative frequency of sv_i in every equivalence class of $\mathcal{E}$, that is $|(e, sv_i)|/|e|$, is no more than α. Where (e, sv_i) be the set of tuples in equivalence class e containing sv_i and α. A limitation of (α, k)-anonymity is that it may result in a high level of distortion if the values for sensitive attributes are skewed [44].

2.2.5 *m*-invariance

The *m*-invariance definition is also built upon k-anonymity. Most of the privacy definitions are suitable for the scenario where the relation is to be released once only. Second or subsequent releases of the relation (with an updated record or insertion of a new record) may lead to inferences.

The *m*-invariance definition [45] states that *"an anonymized relation $\mathcal{T}$ is m-unique if each equivalence class $e \in \mathcal{T}$ contains at least m set of records (or tuples) and all the records in e must have different sensitive attribute values. The sequence of published relations is said to be m-invariant if all the releases are m-unique, along with the condition that the set of sensitive attribute values for every e in $\mathcal{T}$ must remain same for subsequent releases of relation $\mathcal{T}$."*

This privacy definition captures dynamic republication of relations. However, the condition that values of the sensitive attribute should remain the same, as are in previous releases, is strong and limits its applicability.

2.2.6 (k, e)-anonymity

The (k, e)-anonymity definition is an alteration of k-anonymity tailored for numeric data and can only be applied to sensitive attributes having numeric values. (k, e)-anonymity requires that the range of the equivalence class e to be larger than a certain threshold [46].

(k, e)-anonymity is susceptible to *proximity attack* [47]. A proximity breach occurs when the adversary predicts a short interval for an individual's numeric sensitive attribute's value but not the exact value for the sensitive value itself.

2.2.7 (ϵ, m)-anonymity

The (ϵ, m)-anonymity definition is a refinement to (k, e)-anonymity to overcome a proximity breach, and similarly, is only applicable to sensitive attributes having numerical values. The (ϵ, m)-anonymity requires that given an equivalence class e, for every sensitive value $sv_i \in e$, then at most $1/m$ of the tuples in e can have sensitive values 'similar' to sv_i. The factor of ϵ quantifies the similarity. The authors list two ways to quantify the numerical similarity between two numerical values of sensitive attributes [47]: two values are similar if their absolute difference is less than the parameter ϵ, i.e. $|sv_i - sv_j| \leq \epsilon$, and relative similarity where sv_i is similar to sv_j if $|sv_i - sv_j| \leq sv_j \cdot \epsilon$.

2.2.8 Multi-relational k-anonymity

Most of the privacy definitions examined above deal with a single relation. Multi-relational k-anonymity defines privacy for cases involving multiple relations by modifying k-anonymity to accommodate multiple relations settings. The multi-relational k-anonymity privacy assumes that there exists a *Person-specific Relation* $\mathcal{T}_{per}$ with respect to a population $\mathcal{U}$ such that $\mathcal{T}_{per}$ has identifiers (primary key attribute or set of attributes) that uniquely correspond to an individual in population $\mathcal{U}$. Given a set of relations $\mathcal{T}_1, \mathcal{T}_2, \ldots, \mathcal{T}_n$, containing foreign keys along with quasi-identifiers as well as sensitive attributes. The notion of multi-relational k-anonymity is to apply k-anonymity at owner level instead of the record level. As there can be many records, spreading over multiple relations, belonging to a single owner. Multi-relational k-anonymity requires that for each record owner r^{id} in the join of all relations, i.e. $\mathcal{T}_{per} \bowtie \mathcal{T}_1 \bowtie \mathcal{T}_2 \bowtie \ldots \bowtie \mathcal{T}_n$, there are k - 1 other record owners with the same quasi-identifier values [48].

2.2.9 δ-disclosure privacy

The *delta*-disclosure privacy definition can be considered a more restrictive version of t-closeness. Given a set of quasi-identifiers $QI \in Attr \setminus SenAttr$ then tuples r_i and r_j are quasi-identifier equivalent if $r_i[QI] = r_j[QI]$. The relation $\mathcal{T}$ can be divided into quasi-identifier equivalence classes as done in k-anonymity. The delta δ-disclosure privacy defines that *"an equivalence class e is δ-disclosure-private with respect to the sensitive attribute in SAV if, for all $sv_i \in SAV$,*

$\left| log \frac{p(e,sv_i)}{p(\mathcal{T},sv_i)} \right| < \delta$. *Where $p(X, sv_i)$ is the probability that a randomly chosen member of X has sensitive attribute value sv_i."* A relation $\mathcal{T}$ is δ-disclosure private if for every $e \in \mathcal{E}$ is δ-disclosure private [49]. However, the δ-disclosure privacy definition is too strong for practical reasons because an equivalence class may not be able to cover all the sensitive attribute values.

2.3 Differential privacy (ϵ-differential Privacy)

Differential privacy requires that any given disclosure is, within a small multiplicative factor that is ϵ, just as likely regardless of whether or not the individual's record in the relation [28]. Intuitively, consider two datasets DS_1 and DS_2, only differing is in one record. Res_1 and Res_2 are the result of query Q to DS_1 and DS_2. Res_1 and Res_2 must be indistinguishable from each other in order to fulfil differential privacy requirement. This is usually achieved by the addition of noise to the query results.

Many differential private algorithms have been proposed [50, 51, 52, 53, 54, 55]. Differentially private algorithms are usually designed for interactive query settings; however, with a limitation that they answer only a limited number of queries – governed by privacy budget. The privacy budget regulates that number of queries after which the answers to the queries no longer considered free from privacy risk. Additionally, in general, differential privacy works for statistical databases. Statistical databases deal with aggregates, in contrast with microdata release. In some scenarios the amount of noise added to the output degrades the utility of the data [56].

Besides the reviewed definitions the privacy literature encompasses many formal privacy definitions and models including integral privacy [57], (X,Y)-privacy [58], and β-likeness [59].

3 Summary

This paper reviewed some of the definitions of anonymity. Observations from the literature are summarized as follows:

- ***k*-anonymity as a foundation:** k-anonymity being one of the first privacy definitions, provided a foundation to further privacy research. Several privacy definitions are based on k-anonymity.
- **Majority of the definitions are for one-time publication:** the privacy literature encompasses a large number of privacy definition, mainly for one-time publication of a single relation [32, 29, 30, 31, 58, 45, 47, 60, 49].
- **Research lacks privacy definitions supporting interactive query settings for microdata release:** the privacy research lacks privacy definitions and mechanisms that work in interactive setting. While differential privacy supports interactive queries, the mechanisms are constrained in the number of queries permitted, and secondly, it allows aggregate queries instead of microdata (row-level data), in some scenarios the amount of noise added degrades the utility of the data.

- **Susceptibility to attacks:** privacy definitions have a level of susceptibility to attacks. There exist vulnerabilities and scenarios where the privacy restriction laid down by the privacy definition is met, but privacy is compromised. Their weaknesses are widely studied in the literature; for instance, the paper from Josep Domingo-Ferrer and Vicenç Torra have discussed the shortcomings of well-known syntactic privacy models in [42]. Therefore, novel ways to detect these privacy attacks are always desirable.
- **Building privacy-preserving interactive query mechanism – a challenge:** in the present era of big data and data analytics, approaches supporting unlimited interactive querying along with permitting the queries to return microdata while preserving privacy is desirable. Therefore, weaker notions of privacy for interactive query settings for microdata release are potential starting point. However, development interactive query settings for microdata release is a challenging task as multiple releases may give rise to inferences being made by an adversary.

4 Conclusions

The privacy literature encompasses a large number of privacy definitions, mainly for one-time publication of a single relation. Most of these definitions are evolved from k-anonymity. k-anonymity and differential privacy can be considered the two mainstream definition of privacy. The privacy research lack privacy definitions and mechanisms that works for interactive settings for microdata release. Privacy definitions, in various scenarios, are susceptible to attacks – where privacy definitions are met, but still the privacy is compromised. Therefore, novel ways to detect these privacy attacks are also desirable.

References

[1] Ferdinand Schoeman. Privacy: Philosophical dimensions. *American Philosophical Quarterly*, 21(3):199–213, 1984.

[2] Michael Zuckerman. Privacy in colonial new england. by david h. flaherty. *Journal of American History*, 59(3):679–681, 1972.

[3] Osborne M. Reynolds. *Administrative Law Review*, 22(1):101–106, 1969.

[4] Polyvios G. Polyviou. *Search & seizure : constitutional and common law Polyvios G. Polyviou*. Duckworth London, 1982.

[5] W. A. Parent. Privacy, morality, and the law. *Philosophy & Public Affairs*, 12(4):269–288, 1983.

[6] Charles Fried. Privacy. *The Yale Law Journal*, 77(3):475–493, 1968.

[7] Louis Henkin. Privacy and autonomy. *Columbia Law Review*, 74(8):1410–1433, 1974.

[8] Ruth Gavison. Privacy and the limits of law. *The Yale Law Journal*, 89(3):421–471, 1980.

[9] Christopher H. Pyle. Privacy, law, and public policy. by david m. obrien. (new york: Praeger publishers, 1979.). *American Political Science Review*, 75(1):206–207, 1981.

[10] *Universal Declaration of Human Rights*. December 1948. Online at: `https://www.ohchr.org/EN/UDHR/Documents/UDHR_Translations/eng.pdf`.

[11] Centers for Medicare & Medicaid Services. The Health Insurance Portability and Accountability Act of 1996 (HIPAA). Online at http://www.cms.hhs.gov/hipaa/, 1996.

[12] 2018 reform of EU data protection rules, 2018. Online at: `https://ec.europa.eu/commission/sites/beta-political/files/data-protection-factsheet-changes_en.pdf`.

[13] OECD. The OECD privacy framework. Technical report, OECD Publishing, 2013. Online at: `https://www.oecd.org/sti/ieconomy/oecd_privacy_framework.pdf`.

[14] X. Yang, C. M. Procopiuc, and D. Srivastava. Recommending join queries via query log analysis. In *2009 IEEE 25th International Conference on Data Engineering*, pages 964–975, March 2009.

[15] T. Dalenius. Towards a methodology for statistical disclosure control. *Statistik Tidskrift*, 15(429-444):2–1, 1977.

[16] Joshua Fogel and Elham Nehmad. Internet social network communities: Risk taking, trust, and privacy concerns. *Computers in Human Behavior*, 25(1):153 – 160, 2009.

[17] Randy Baden, Adam Bender, Neil Spring, Bobby Bhattacharjee, and Daniel Starin. Persona: An online social network with user-defined privacy. In *Proceedings of the ACM SIGCOMM 2009 Conference on Data Communication*, SIGCOMM '09, pages 135–146, New York, NY, USA, 2009. ACM.

[18] L. A. Cutillo, R. Molva, and T. Strufe. Safebook: A privacy-preserving online social network leveraging on real-life trust. *IEEE Communications Magazine*, 47(12):94–101, Dec 2009.

[19] S. Sicari, A. Rizzardi, L.A. Grieco, and A. Coen-Porisini. Security, privacy and trust in internet of things: The road ahead. *Computer Networks*, 76:146 – 164, 2015.

[20] A. Ukil, S. Bandyopadhyay, and A. Pal. Iot-privacy: To be private or not to be private. In *2014 IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS)*, pages 123–124, April 2014.

[21] P. Porambage, M. Ylianttila, C. Schmitt, P. Kumar, A. Gurtov, and A. V. Vasilakos. The quest for privacy in the internet of things. *IEEE Cloud Computing*, 3(2):36–45, Mar.-Apr. 2016.

[22] Anil Gurung and M.K. Raja. Online privacy and security concerns of consumers. *Information and Computer Security*, 24(4):348–371, 2016.

[23] George R. Milne, Andrew J. Rohm, and Shalini Bahl. Consumers' protection of online privacy and identity. *The Journal of Consumer Affairs*, 38(2):217–232, 2004.

[24] Giannakis Antoniou and Lynn Batten. E-commerce: protecting purchaser privacy to enforce trust. *Electronic Commerce Research*, 11(4):421, Aug 2011.

[25] E. F. Codd. A relational model of data for large shared data banks. *Commun. ACM*, 13(6):377–387, 1970.

[26] Adrienne Watt and Nelson Eng. *Database Design*. BCcampus, 2014. Online at: `https://opentextbc.ca/dbdesign01/`.

[27] Syed Rafiul Hussain, Asmaa M. Sallam, and Elisa Bertino. Detanom: Detecting anomalous database transactions by insiders. In *Proceedings of the 5th ACM Conference on Data and Application Security and Privacy*, CODASPY '15, pages 25–35, New York, NY, USA, 2015. ACM.

[28] Cynthia Dwork. Differential privacy: A survey of results. In Manindra Agrawal, Dingzhu Du, Zhenhua Duan, and Angsheng Li, editors, *Theory and Applications of Models of Computation*, pages 1–19, Berlin, Heidelberg, 2008. Springer Berlin Heidelberg.

[29] Ashwin Machanavajjhala, Daniel Kifer, Johannes Gehrke, and Muthuramakrishnan Venkitasubramaniam. L-diversity: Privacy beyond k-anonymity. *ACM Trans. Knowl. Discov. Data*, 1(1), March 2007.

[30] N. Li, T. Li, and S. Venkatasubramanian. t-closeness: Privacy beyond k-anonymity and l-diversity. In *2007 IEEE 23rd International Conference on Data Engineering*, pages 106–115, April 2007.

[31] Raymond Chi-Wing Wong, Jiuyong Li, Ada Wai-Chee Fu, and Ke Wang. (α, k)-anonymity: An enhanced k-anonymity model for privacy preserving data publishing. In *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '06, pages 754–759, New York, NY, USA, 2006. ACM.

[32] Latanya Sweeney. k-anonymity: A model for protecting privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 10(5):557–570, 2002.

[33] Latanya Sweeney. Simple demographics often identify people uniquely. Working paper, 2000. Working paper. Online at: `http://dataprivacylab.org/projects/identifiability/`.

[34] P. Zhao, J. Li, F. Zeng, F. Xiao, C. Wang, and H. Jiang. Illia: Enabling*k*-anonymity-based privacy preserving against location injection attacks in continuous lbs queries. *IEEE Internet of Things Journal*, 5(2):1033–1042, April 2018.

[35] Yu-Meng Ye, Chang-Chun Pan, and Gen-Ke Yang. An improved location-based service authentication algorithm with personalized k-anonymity. In Jiadong Sun, Jingnan Liu, Shiwei Fan, and Feixue Wang, editors, *China Satellite Navigation Conference (CSNC) 2016 Proceedings: Volume I*, pages 257–266, Singapore, 2016. Springer Singapore.

[36] Y. Zhang, W. Tong, and S. Zhong. On designing satisfaction-ratio-aware truthful incentive mechanisms for$k$-anonymity location privacy. *IEEE Transactions on Information Forensics and Security*, 11(11):2528–2541, Nov 2016.

[37] Yingjie Wang, Zhipeng Cai, Zhongyang Chi, Xiangrong Tong, and Lijie Li. A differentially k-anonymity-based location privacy-preserving for mobile crowdsourcing systems. In Rongfang Bie, Yunchuan Sun, and Jiguo Yu, editors, *2017 International Conference on Identification, Information and Knowledge in the Internet of Things, IIKI 2017, Shandong, China, October 19-21, 2017*, volume 129 of *Procedia Computer Science*, pages 28–34. Elsevier, 2017.

[38] Sheng Zhong, Hong Zhong, Xinyi Huang, Panlong Yang, Jin Shi, Lei Xie, and Kun Wang. *Connecting Things to Things in Physical-World: Security and Privacy Issues in Vehicular Ad-hoc Networks*, pages 101–134. Springer International Publishing, Cham, 2019.

[39] Y. Khazbak, J. Fan, S. Zhu, and G. Cao. Preserving location privacy in ride-hailing service. In *2018 IEEE Conference on Communications and Network Security (CNS)*, pages 1–9, May 2018.

[40] Dotan Di Castro, Liane Lewin-Eytan, Yoelle Maarek, Ran Wolff, and Eyal Zohar. Enforcing k-anonymity in web mail auditing. In *Proceedings of the Ninth ACM International Conference on Web Search and Data Mining*, WSDM '16, pages 327–336, New York, NY, USA, 2016. ACM.

[41] Junade Ali. Mechanism for the prevention of password reuse through anonymized hashes. *PeerJ PrePrints*, 5:e3322, 2017.

[42] J. Domingo-Ferrer and V. Torra. A critique of k-anonymity and some of its enhancements. In *2008 Third International Conference on Availability, Reliability and Security*, pages 990–993, March 2008.

[43] Hongyu Liang and Hao Yuan. On the complexity of t-closeness anonymization and related problems. In Weiyi Meng, Ling Feng, Stéphane Bressan, Werner Winiwarter, and Wei Song, editors, *Database Systems for Advanced Applications*, pages 331–345, Berlin, Heidelberg, 2013. Springer Berlin Heidelberg.

[44] Benjamin C.M. Fung, Ke Wang, Ada Wai-Chee Fu, and Philip S. Yu. *Introduction to Privacy-Preserving Data Publishing: Concepts and Techniques*. Chapman & Hall/CRC, 1st edition, 2010.

[45] Xiaokui Xiao and Yufei Tao. M-invariance: Towards privacy preserving re-publication of dynamic datasets. In *Proceedings of the 2007 ACM SIGMOD International Conference on Management of Data*, SIGMOD '07, pages 689–700, New York, NY, USA, 2007. ACM.

[46] Q. Zhang, N. Koudas, D. Srivastava, and T. Yu. Aggregate query answering on anonymized tables. In *2007 IEEE 23rd International Conference on Data Engineering*, pages 116–125, April 2007.

[47] Jiexing Li, Yufei Tao, and Xiaokui Xiao. Preservation of proximity privacy in publishing numerical sensitive data. In *Proceedings of the 2008 ACM SIGMOD International Conference on Management of Data*, SIGMOD '08, pages 473–486, New York, NY, USA, 2008. ACM.

[48] M. E. Nergiz, C. Clifton, and A. E. Nergiz. Multirelational k-anonymity. *IEEE Transactions on Knowledge and Data Engineering*, 21(8):1104–1117, Aug 2009.

[49] Justin Brickell and Vitaly Shmatikov. The cost of privacy: destruction of data-mining utility in anonymized data publishing. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Las Vegas, Nevada, USA, August 24-27, 2008*, pages 70–78, 2008.

[50] Q. Geng and P. Viswanath. The optimal mechanism in differential privacy. In *2014 IEEE International Symposium on Information Theory*, pages 2371–2375, June 2014.

[51] F. McSherry and K. Talwar. Mechanism design via differential privacy. In *48th Annual IEEE Symposium on Foundations of Computer Science (FOCS'07)*, pages 94–103, Oct 2007.

[52] Chao Li, Michael Hay, Vibhor Rastogi, Gerome Miklau, and Andrew McGregor. Optimizing linear counting queries under differential privacy. In *Proceedings of the Twenty-ninth ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems*, PODS '10, pages 123–134, New York, NY, USA, 2010. ACM.

[53] Moritz Hardt, Katrina Ligett, and Frank McSherry. A simple and practical algorithm for differentially private data release. *CoRR*, abs/1012.4763, 2010.

[54] Jianping He and Lin Cai. Differential private noise adding mechanism: Fundamental theory and its application. *CoRR*, abs/1611.08936, 2016.

[55] Cynthia Dwork and Aaron Roth. The algorithmic foundations of differential privacy. *Found. Trends Theor. Comput. Sci.*, 9(3–4):211–407, August 2014.

[56] Simson L. Garfinkel, John M. Abowd, and Sarah Powazek. Issues encountered deploying differential privacy. In *Proceedings of the 2018 Workshop on Privacy in the Electronic Society*, WPES'18, pages 133–137, New York, NY, USA, 2018. ACM.

[57] Julien Aligon, Matteo Golfarelli, Patrick Marcel, Stefano Rizzi, and Elisa Turricchia. Similarity measures for olap sessions. *Knowledge and Information Systems*, 39(2):463–489, May 2014.

[58] Ke Wang and Benjamin C. M. Fung. Anonymizing sequential releases. In *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '06, pages 414–423, New York, NY, USA, 2006. ACM.

[59] Jianneng Cao and Panagiotis Karras. Publishing microdata with a robust privacy guarantee. *PVLDB*, 5(11):1388–1399, 2012.

[60] Mehmet Ercan Nergiz, Maurizio Atzori, and Chris Clifton. Hiding the presence of individuals from shared databases. In *Proceedings of the 2007 ACM SIGMOD International Conference on Management of Data*, SIGMOD '07, pages 665–676, New York, NY, USA, 2007. ACM.