

Characterising and Categorising Anonymization Techniques: A Literature-Based Approach

Andrea Fieschi^{1,2}^a, Pascal Hirmer¹^b, Christoph Stach²^c and Bernhard Mitschang²^d

¹Mercedes-Benz AG, Stuttgart, Germany

²Institute for Parallel and Distributed Systems, University of Stuttgart, Stuttgart, Germany

{andrea.fieschi, pascal.hirmer}@mercedes-benz.com,
{ch .de

Keywords: Privacy Protection, PRISMA Systematic Literature Research, Privacy-Enhancing Techniques, Anonymization Techniques.

Abstract: Anonymization plays a crucial role in protecting personal data and ensuring information security. However, selecting the appropriate anonymization technique is a challenging task for developers, data scientists, and security practitioners due to the vast array of techniques available in both research and practice. This paper aims to assist users by offering a method for structuring a framework that helps them make informed decisions about the most appropriate anonymization techniques for their specific use cases. To achieve this, we first conduct a systematic literature review following the PRISMA guidelines to capture the current state of the art in anonymization techniques. Based on the findings from this review, we propose a conceptual organisation of anonymization techniques, designed to help users navigate the complex landscape of anonymization and choose techniques that align with their security requirements.

1 INTRODUCTION


Data collection is a necessity in various domains, but it poses significant risks to information security. As the volume of data collected from sources like patients, smartphones, or vehicles increases, so does the potential for exposing sensitive information that individuals did not agree to disclose. In this context, protecting personal privacy and securing data are critical challenges in information security (Stach, 2023).


In order to protect personal privacy, numerous Privacy Enhancing Technologies (PETs) can be used. A possible way of achieving effective privacy protection is anonymization (Majeed and Lee, 2021). The European General Data Protection Regulation (GDPR) defines anonymization as an "irreversible transformation of personal data in such a way that the data subject can no longer be identified" (European Parliament and Council of the European Union, 2016).


Following the concept of Anonymization by Design (Fieschi et al., 2024), choosing the most suit-


able anonymization technique early in the development stages of a data-collecting use case is essential for ensuring effective privacy protection. Privacy can guarantee stronger protection if it is considered during the development process (Morton and Sasse, 2012). However, selecting the right anonymization technique for a particular data use case presents a significant challenge for developers, data scientists, and security practitioners. Privacy needs to cater to the needs of both the service provider and the service users (Fieschi et al., 2023). The heterogeneous landscape of anonymization techniques, each with its own strengths and limitations, makes it difficult to choose the most suitable approach, especially when aiming to ensure robust information security. The lack of a comprehensive, organised framework further complicates the decision-making process, increasing the risk of inappropriate or ineffective techniques being used, potentially leading to security breaches.

This paper addresses this gap by proposing how to structure a collection of anonymization techniques that offers a comprehensive overview, designed to support users (developers, data scientists, and security practitioners) in selecting the appropriate technique to meet their use case security requirements. Our goal is to organise a conceptual framework that as-

^a <https://orcid.org/0009-0007-9126-6021>

^b <https://orcid.org/0000-0002-2656-0095>

^c <https://orcid.org/0000-0003-3795-7909>

^d <https://orcid.org/0000-0003-0809-9159>

sists security-conscious users in navigating the complex landscape of anonymization techniques and making informed decisions that ensure both privacy and security. To this end, it is important that the framework supporting users provides an easy-to-navigate and thorough overview of the available anonymization techniques; that it's not monolithic and allows each technique to be used as a single piece; and that it is flexible enough to allow users to incorporate new techniques published in the literature, developed in practice, or self-developed.

In this paper, we provide two main contributions. First, we present the insights gained about the landscape of anonymization techniques through a systematic literature review, conducted using the PRISMA method (Moher et al., 2015). It was important for us to understand the types of techniques available, the kinds of data they process, and the domains in which they are applied. Second, we propose a conceptual structure for organising anonymization techniques, based on our literature research findings.

In Section 2, we present the current state of the art of collections of anonymization techniques and we highlight the reasons why the present solutions do not fully satisfy our needs. Section 3 details our PRISMA-based literature review and characterises the anonymization techniques landscape. Building on these insights, Section 4 explains how our findings inform the structured organisation of a new collection of anonymization techniques. We then compare our proposed framework with existing solutions in Section 5, highlighting its advantages and addressing current shortcomings. Finally, Section 7 concludes the paper by summarising the key takeaways and suggesting directions for future research and development.

2 RELATED WORK

In the field of data anonymization, ensuring robust privacy protection requires careful consideration of available techniques. To identify the most suitable anonymization technique, security-conscious users would benefit from having access to a well-organised collection of available techniques that will: 1) *Provide a Comprehensive Overview*: Enable users to make well-informed decisions by offering a complete range of anonymization techniques for various use cases. 2) *Offer Modular Deployment*: Design each technique as an individual module so that users can deploy only the specific technique they need, rather than integrating an entire framework. This approach simplifies deployment and minimises overhead. 3) *Offer Up-datability*: Allow users to add new techniques as they

emerge from research, or are self-developed in order to maintain the collection relevant and current.

Several software solutions have been created to offer a range of anonymization techniques and support their implementation. However, these solutions come with notable limitations that can compromise their effectiveness in supporting the selection of the most suitable anonymization technique in a real-world scenario. The solutions found in the practice bring together only anonymization techniques of a similar nature and act on a focused part of a data stream.

There are software solutions that provide privacy protection through anonymization by acting at the very beginning of the data stream. Before the data are used in any way, these methodologies generate a new dataset with the same characteristics as the one acquired from real-world scenarios. The Synthetic Data Vault (Patki et al., 2016) and the Synthetic Data Generation framework (Walonoski et al., 2017) are examples of frameworks that provide methodologies for generating synthetic datasets which can be used for testing and development without risking real data exposure. These tools are valuable for creating safe environments for data analysis, but they are not directly focused on the anonymization of existing datasets.

Other software solutions bring together anonymization techniques that are apt at modifying the acquired dataset before analysing it or passing it on to the next data handler. ARX (Prasser and Kohlmayer, 2015) is a good example as it is a notable software that provides a rich set of anonymization techniques, including k-anonymity, l-diversity, and t-closeness. It is a well-regarded tool in the field, appreciated for its user-friendly interface and robust algorithmic implementations. However, despite its strengths, ARX is not without limitations. Specifically, it falls short in its integration capabilities; once an anonymization technique is selected, deploying it within a data processing pipeline is not straightforward. Its monolithic nature does not allow for a single algorithm to be used. ARX is designed more as a standalone application rather than a modular component that can be seamlessly incorporated into existing workflows. OpenDP (Gaboardi et al., 2020) is another good example, some of its anonymization techniques consist in adding noise according to the Differential Privacy (DP) postulate, hence modifying the dataset before it is sent to the next stage of the data stream. It too comes with the same limitations mentioned for ARX.

There are also software solutions that offer privacy protection by employing anonymization techniques that alter the way data is handled for its intended use. OpenDP (Gaboardi et al., 2020) and TensorFlow Pri-

vacy (Abadi et al., 2015) are good examples of it since they offer an ensemble of techniques apt for ensuring differential privacy. The first provides mechanisms that allow differentially private queries, while the latter ensures the successful employment of differential privacy within machine learning models. While these libraries are powerful in their respective niches, they do not provide an overview of all the available anonymization techniques. Moreover, they too are not designed as modular, making them ill-suited for on-the-fly deployment.

All of the above solutions offer a range of methods, though often not as comprehensive as one might desire for a flexible and modular approach. These software solutions present the following problems: 1) *Fragmented Coverage*: The focus of each one of them on a specific type of anonymization approach does not allow us to use any of these software solutions as a comprehensive overview of all the different techniques available. This fragmentation can obstruct users from making well-informed decisions based on a full range of options. 2) *Integration Issues*: The standalone nature of many solutions complicates their integration into existing data processing workflows. This lack of modularity limits the practical applicability of these tools in dynamic and evolving data environments. 3) *Limited Flexibility*: Existing solutions often fail to incorporate new techniques emerging from recent research, address user-specific needs, or allow users to incorporate custom techniques, resulting in a lack of adaptability and relevance.

We need to lay the groundwork that allows us to organise a collection of anonymization techniques with the characteristics listed at the beginning of this section. The first step to this end is conducting a systematic literature review, which will serve as the foundation for developing our proposed collection.

3 LITERATURE ANALYSIS: ANONYMIZATION TECHNIQUES

To build an organised collection of anonymization techniques, it is important to get a clear picture of the methods that exist in the literature. To this end, we conducted a systematic literature review using the PRISMA method (Moher et al., 2015). By reviewing the anonymization techniques available in the literature, we were able to determine how to structure our collection of these methods.

Notable surveys on the topic like (Chen et al., 2009), provide a comprehensive explanation of the

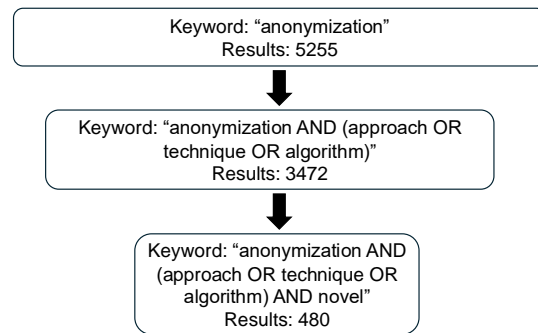


Figure 1: Keywords Identification for the Papers Research.

subject and its various approaches. Our PRISMA research provides us with insights into the literature landscape of anonymization.

3.1 PRISMA Research

The PRISMA method (Moher et al., 2015), short for Preferred Reporting Items for Systematic Reviews and Meta-Analyses, is a protocol that guides researchers through the process of conducting a systematic literature review. The approach helps ensure comprehensive coverage and an unbiased selection of relevant studies. It begins with defining strict criteria for which studies to include and exclude. Researchers then search relevant databases and sources using a detailed strategy, followed by a careful screening of studies based on titles, abstracts, and full texts to ensure relevance. Data is extracted from the chosen studies using a standardised approach, and the quality of each study is assessed to identify any potential bias. The results from these studies are then synthesised, either quantitatively or qualitatively, to form conclusions. Finally, the process and findings are reported in a structured and transparent way, often accompanied by a flow diagram to map out the study selection process. This methodical approach is designed to ensure clarity, thoroughness, and reproducibility in the review of research literature.

Our research was conducted in 2023 consulting the Scopus database¹ and web of science². Both online platforms index and store peer-reviewed literature from various sources such as IEEE, Elsevier, and Springer. The user-friendly interfaces of both platforms enable efficient refinement of search results according to various criteria. Since both Scopus and Web of Science collect scientific works from more or less the same sources, the results yielded were approximately the same for this research. Therefore, we

¹<https://www.scopus.com/>

²<https://www.webofscience.com/>

chose to use only one of the two platforms and our choice fell on Scopus.

In order to understand the number of papers present in literature, we started by entering only the word "anonymization" as a search term. We included both the British and American spellings of the word. For practicality, we will only use the American spelling in the keyword list. This yielded 5255 results. Given the size of the vast amount of papers, we proceeded with further refining the search by adding more search terms. In Figure 1, we see how the number of results changes when we refine the keyword search. To focus more, we refined the search term with "anonymization AND (approach OR technique OR algorithm)", this reduces the number of results to 3472. The addition of "novel" to the research terms helped us weed out papers that marginally talk about anonymization or that are not proposing a new approach to an anonymization technique.

In Figure 2, we see the PRISMA scheme flow that led us to identify the papers to be analysed and included in our search. Through stages of search refinement and early screening, we managed to eliminate the papers that would not have contributed to our literature search by following the PRISMA paradigm. With 139 papers we have a reasonable reflection of the works present in the literature and the statistical information we extrapolate from this ensemble helps us gain a clear overview of the kind of anonymization techniques present in the literature, their application domains, and the type of data types processed. In Section 3.2, we see the main information we extrapolated from these results.

3.2 Anonymization Categories

In the literature, we found several different anonymization techniques. The multitude of approaches can appear rather overwhelming. However, a pattern can be traced among all of them. To put some order in the landscape of anonymization techniques we defined 5 overarching categories, as it can be seen in Section 3.3, under which most techniques can be grouped. In the following, we explain the categories we identified, give them a name, and reference the main techniques belonging to each category. It has to be noted that the basic step common to all anonymization techniques is eliminating the direct identifiers, i.e., attributes that directly link the data to a specific data source, such as full names, ID numbers, matriculation numbers, etc. This is not enough to guarantee anonymity since the collection of further attributes that describe the data source, i.e., quasi-identifier, can still lead to the risk of identifying the data source.

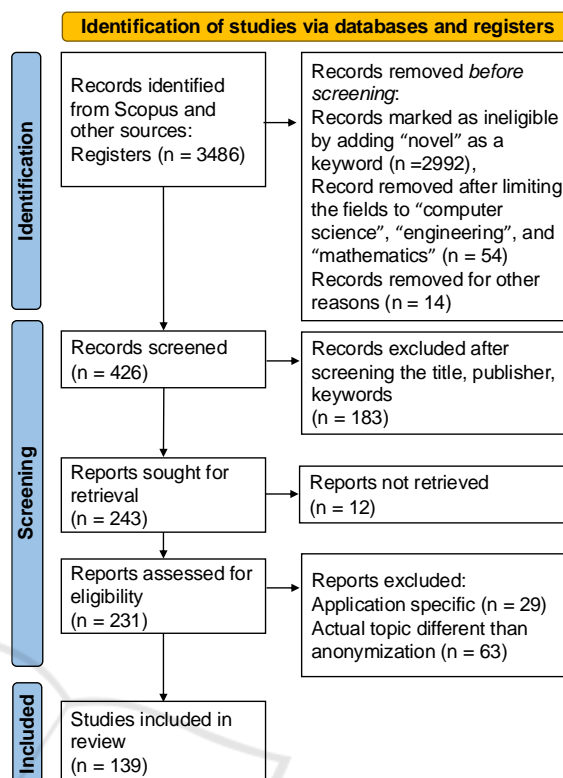


Figure 2: PRISMA Flowchart.

Therefore, for all the categories we describe in the following sections, this step is included.

3.2.1 Grouping Based

With the term "grouping based", we refer to anonymization techniques that try to guarantee anonymity

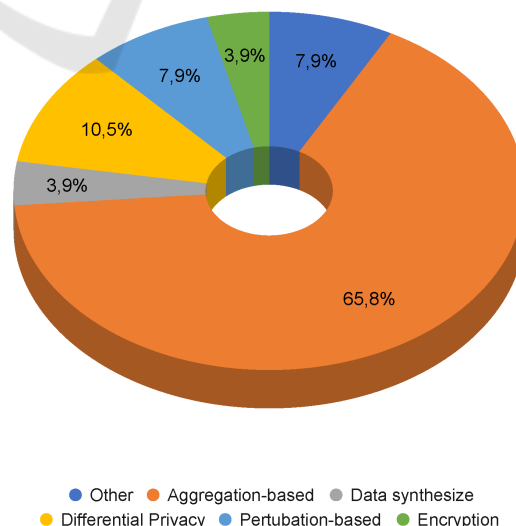


Figure 3: Visualisation of the distribution of papers in the literature according to our categorisation.

Table 1: Anonymization techniques in the literature.

Anonymization Model	Explanation	Examples of Sources
Grouping Based	k -anonymity, l -diversity, t -closeness, etc.	(Sweeney, 2002), (Li et al., 2007), (Machanavajjhala et al., 2007), (Khan et al., 2022)
Differential Privacy	Reaching the guarantee of the DP postulate	(Dwork, 2008), (Dwork and Roth, 2013), (Yu et al., 2019)
Perturbation Based	Obfuscation through data perturbation, e.g., noise injection	(Hamm, 2017), (Aljably, 2021), (Shynu et al., 2020)
Data Synthesize	Generate fake data with the same properties of real data	(Piacentino and Angulo, 2020a), (Piacentino and Angulo, 2020b)
Encryption	Used to enhance anonymity protection, e.g., block-chain.	(Javed et al., 2021), (Alnemari et al., 2018), (Yamaç et al., 2019)
Other	Specific techniques for specific data types, like voice or video	(Fan and Wang, 2023), (Zhao et al., 2016)

by creating groups of data sources that are indistinguishable from one another. This is done by modifying the quasi-identifier to allow data coming from different sources to be grouped and to become indistinguishable from one another. The sensitive attributes are not modified, and can still provide a good level of data usability, but the value of a sensitive attribute under protection will not be re-linked to its source since it is grouped with data coming from multiple sources. In most cases, we found papers that implement or extend k -anonymity (Sweeney, 2002), l -diversity (Machanavajjhala et al., 2007), or t -closeness (Li et al., 2007).

3.2.2 Differential Privacy

This cluster contains all the papers that present anonymization techniques that aim at guaranteeing the postulate of DP (Dwork, 2008), that is:

$$Pr[M(D_1) \in S] \leq e^\epsilon Pr[M(D_2) \in S]$$

In the literature (Zhu et al., 2017), the parameter ϵ is also referred to with the term *privacy budget*. In order to achieve higher privacy guarantees, we need a lower value of ϵ . DP can be achieved through differentially private queries (Dwork and Roth, 2013). It could also be reached through randomised response or a mechanism of differentially private data collection (Wang et al., 2016). Most papers aim at improving data usability in different environments or to specific data types as we discovered in (Jin et al., 2022), (Hamm, 2017), (Aljably, 2021), or (Gao and Li, 2019c). While other anonymization approaches can have similar mechanisms, e.g., noise injection, only the approaches that aim at guaranteeing the DP postulate mentioned above are found in this category.

3.2.3 Perturbation Based

Here, we cluster all the anonymization techniques that aim to ensure anonymity by adding noise to the data

or perturbing their values in other ways. These techniques do not aim to satisfy specific postulates like DP or k -anonymity, which is why they are grouped into a separate category. Here, we have anonymization techniques that use value perturbation but do not fall under the category of DP or grouping-based methods. This can be applied at different stages of the data pipeline (Chen et al., 2009). For example, sampling noise from a normal distribution and adding it to an attribute (Domingo-Ferrer et al., 2020).

It has to be pointed out that many anonymization techniques clustered under the Differential Privacy bubble also introduce noise in order to reach the DP postulate. Therefore, they are stricter and have a mathematically defined goal. Some examples are: (Sun et al., 2016), (Ullah and Shah, 2016), (Eyupoglu et al., 2018a), and (Attallah et al., 2021).

3.2.4 Data Synthesis

Under the category Data Synthesis, we have the techniques that protect the users' privacy by synthesising completely new data based on the original data (Fung et al., 2010). With this type of approach, the direct identifiers are often removed before synthesis or pseudonymized in order not to store unprotected data. After the generation, the new dataset is made of data points with no connections to specific users, and the original dataset is deleted. Hence, the guarantee of anonymity. Generative Adversarial Networks can be used to generate new datasets with the same characteristics as the original dataset but are not connected to any real data source (Park et al., 2018). Further examples can be found here: (Aleroud et al., 2022), (Abay et al., 2019), (Piacentino and Angulo, 2020b).

3.2.5 Encryption

The cluster of encryption is an interesting one under the anonymization lens. Encryption is, strictly speaking, not an anonymization technique, however, some techniques use encryption as a base to reach a guar-

antee of anonymity. This can be done in combination with other methods, to obfuscate sensitive data, or to use blockchain methods. The following resources provide valid examples of which kind of technique can converge to this cluster: (Javed et al., 2021), (Alnemari et al., 2018), and (Yamaç et al., 2019).

3.3 Research Landscape: Methods, Application Domains, and Data Types

The systematic literature search helped us understand which anonymization techniques are present in the literature, in which domains they are applied, and which data types they can process. In the following, we illustrate the statistics of which percentage of the papers analysed in our literature can be grouped under a certain category, in which field they are applied, and which data types can be processed by the anonymization techniques they present.

3.3.1 Categories of Anonymization Techniques

In Section 3.2, we outlined the categories of techniques we defined to group the techniques we found in the literature, explaining the nomenclature and what belongs to each category. Here, we illustrate how the works analysed in our literature search are distributed over all the different clusters. Table 1 shows the distribution of all the techniques found. *Grouping-based* techniques are most prevalent, used in about 66% of the papers, often involving k -anonymity and its extensions, i.e., l -diversity and t -closeness.

The next significant group, around 10%, employs *DP*. Different approaches are proposed in the various papers to try to improve data usability for different application fields. *Perturbation-based* techniques make up 8% of the works we analysed, outlining various ways of injecting noise, or generally perturbing the data, in order to achieve anonymization guarantees. Techniques that can be ordered under the cluster of *data synthesis* are found in 4% of the analysed papers. *Encryption-related* anonymization techniques make up 4% of this literature search. As already mentioned in Section 3.2.5, this is peculiar given that encryption is not, strictly speaking, an anonymization method. However, the paper mentioned here uses encryption to strengthen the guarantee of anonymization. The last category, in Table 1 named, contains all the techniques that do not belong to any of the categories described before. Most of the papers grouped in *Others* deal with specific cases and address privacy problems specific to certain data collections. Some examples are image anonymization, speech anonym-

ization, and video anonymization.

3.3.2 Areas of Application

The application areas, combined with the techniques, of the analysed works are detailed in Figure 4. By application areas, we refer to those mentioned in each paper, not to potential areas where the techniques could be applied. It has to be noted that around 40% of the papers are not tied to a specific area of application and mainly address theoretical aspects, such as introducing new algorithms or improving upon already well-established methods.

Social networks and healthcare are the predominant application areas, with the former including about 20% of the works and the latter including about 15% of the analysed works. The healthcare sector has a history of anonymization research due to clinical study evaluations, with ongoing work to refine these methods (Abbasi and Mohammadi, 2022), (Aminifar et al., 2021)). Social networks are also a key area because of the amount of personal information held that industries want to analyse while guaranteeing privacy protection. Here we find a high proportion of differential privacy (Gao and Li, 2019a), (Gao and Li, 2019c)) and perturbation-based techniques (Al-Kharji et al., 2018), (Rong et al., 2018).

The areas of Big Data and Cloud & Web favour the perturbation-based methods (Eyupoglu et al., 2018b), (Kalia et al., 2021). In all other fields, grouping-based techniques are the most used and discussed in the literature.

3.3.3 Data Types Processed

The analysed papers were closely reviewed for the types of data requiring anonymization, as depicted in Figure 5. This figure illustrates the data types and their associated anonymization methods as found in the analysed papers.

Half of the studies focus on tabular data, which is not surprising given its prevalence and ease of anonymization. About 15% of the studies address anonymizing Graph Data (Thouvenot et al., 2020), (Gao and Li, 2019b), graph data often linked to social networks. Another 15% is dedicated to Positional Data. These data types prompted the development of specialised techniques due to their peculiar structure.

Positional data is split into points of interest (An et al., 2018), (Li et al., 2021), (Sei and Ohsuga, 2017) and trajectory (Ward et al., 2017), (Mahdavi-far et al., 2022), (Li et al., 2022), with the former requiring complex anonymization techniques due to the sensitivity of location data. Streaming and transactional data are less commonly studied, with existing

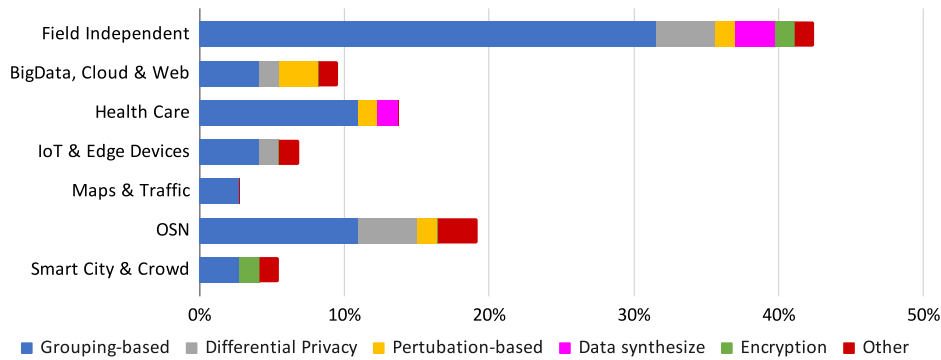


Figure 4: Techniques found in the literature grouped by application domain.

techniques adapted to their specific needs (Mohamed et al., 2020), (Tsai et al., 2020), (Puri et al., 2022). Image, video, and speech data anonymization are not extensively covered in this review, as they require use-case-specific approaches.

3.4 Literature Gaps

As we have illustrated in Figure 5, most of the anonymization methods found in the literature deal with tabular data. Images, videos, and speech data, just to name a few, are handled with methods that are focused on that kind of data type used for a specific use case. Also log data, already closer to tabular data, is a data type not extensively considered in the literature, only few specific cases deal with this data type. An approach that tackles log data would help anonymization being used as a privacy-protecting mechanism for diagnostics or product improvement.

Another aspect lacking in the literature is the guarantee of anonymity also after incremental data updates. Except for a few cases like (Pei et al., 2007), (Dwork and Roth, 2013), most of the anonymization approaches are not run in real-time and in order to maintain the same guarantee they need to be re-run when new data come in. This is particularly true for tabular data and for grouping-based methods. Anonymization thought for incremental data updates would make its use easier for stream data and its application more widespread.

4 HOW TO ORGANISE A COLLECTION OF ANONYMIZATION TECHNIQUES

Leveraging the information extrapolated in section 3, the knowledge from the literature research and the classification we made of the anonymization techniques, we illustrate here our vision on how to organise a collection of anonymization techniques.

The extensive amount of anonymization algorithms necessitates a clear organisational method to exploit their different characteristics, areas of application, and best usages. To achieve this, we employ a hierarchical structure to categorise these techniques systematically. We propose a three-tier hierarchical model to organise anonymization techniques. We devise the following three structuring criteria: 1) *Anonymization Category*: This is the broadest classification level, grouping techniques based on their fundamental approach (e.g., grouping-based, data synthesis, etc.). 2) *Anonymization Technique*: Within each category, specific techniques are detailed (e.g., k-anonymity, differentially private queries). 3) *Anonymization Implementation*: The most granular level, where particular implementations of techniques are described. This includes detailed information about how each technique is applied.

Each level in the hierarchy is modelled with essential attributes: 1) *ID and Name*: Unique identifiers and descriptive names. 2) *Conceptual Explanation*: An overview building on the parent node's explanation. 3) *Data Types*: Types of data the technique can handle. 4) *Application Platform*: Whether it is for backend or on-board processing. 5) *Incremental Updates*: Whether the technique supports updating datasets incrementally. 6) *Implementation De-*

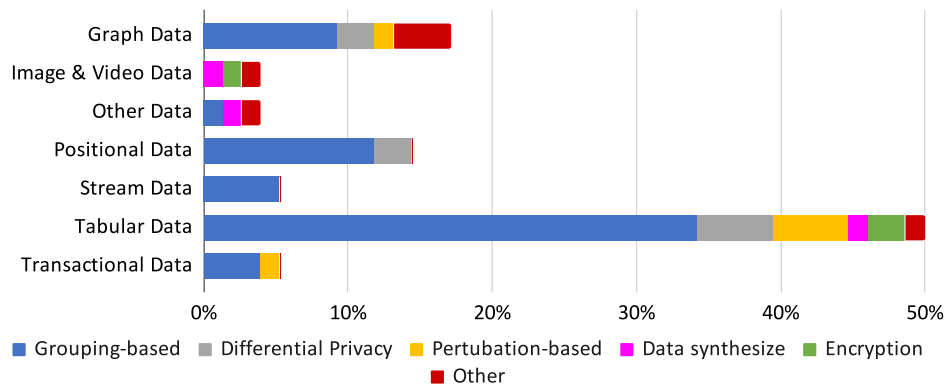


Figure 5: Techniques found in the literature grouped by data type they are applied on.

tails: Specifics on how the technique is implemented.

This hierarchical model, as illustrated in Figure 6 and Figure 7, ensures that each technique and its implementation are well-documented and systematically categorised, hence giving a better overview of the available anonymization techniques. It facilitates modularity by allowing each technique implementation to be treated as separate modules. This modularity is crucial for practical deployment, as it enables users to select and integrate only the specific technique required for their use case, rather than being constrained by a monolithic system.

All the characteristics of a new node are inherited from the parent node, and every attribute can be further specified. For example, the data types that the anonymization technique can process can be limited compared to the parent anonymization category. The same can happen between anonymization technique to the anonymization implementation node.

Every entry of the collection of anonymization techniques is a separate module. Every module follows the model illustrated in Figure 7 of which we can find a description of each attribute in Table 2. The modules from a lower level of the collection of anonymization techniques inherit the characteristics from the level above and add information. On the third level of our hierarchy, we find the techniques' algorithms. The anonymization techniques are here implemented and deployable. The modules of the third level, the anonymization implementation leaf nodes, contain in the description the details of how the anonymity guarantee is reached, through data processing, data pipeline architecture, providing privacy-protecting querying mechanisms, etc. When an anonymization technique can be written as a self-contained and deployable piece of code, e.g., *k*-anonymity, then the implementation is found in the anonymization implementation along with its

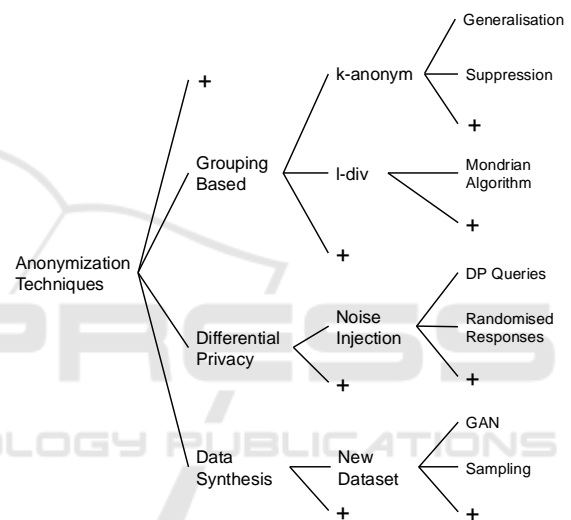


Figure 6: Example of hierarchy and modelling of the collection of anonymization techniques.

documentation, i.e., input and output format, etc. The deployable algorithm could be implemented as a software library, Docker container, binary file, WebAssembly, etc.

Once an anonymization technique is selected for a data-handling process the rest of the collection will not be used in the implementation of the required data pipeline reducing an unrequired overhead otherwise needed by a monolithic structure.

5 COMPARISON OF OUR STRUCTURE WITH ALREADY EXISTING FRAMEWORKS

In this section, we compare our proposed way of organising a collection of anonymization techniques with the state of the art discussed in Section 2. Our goal is to illustrate how our approach distinguishes itself by fulfilling the requirements of providing: a thorough overview, modularity, and flexibility.

A significant advantage of our collection is its modular architecture. The anonymization frameworks mentioned in Section 2 operate as monolithic systems, requiring users to adopt the entire framework even if only a single technique is needed. This can lead to inefficiencies and added complexity. In contrast, our collection is organised and structured with modularity in mind, treating each anonymization technique as an independent module. This allows users to select and deploy only the techniques required for their specific use cases.

Our approach also provides a comprehensive overview of all available anonymization techniques from both literature and practice. Unlike existing frameworks, which may offer a limited or predefined set of techniques, our collection comprises a wide range of methods, ensuring that users have access to the full spectrum of available options. This extensive coverage allows for a more informed selection process, where users can choose the most suitable technique based on their specific needs and use cases. By presenting a complete and organised view of the available techniques, our collection enables users to make well-informed decisions and apply the most effective anonymization methods for their scenarios.

Moreover, current anonymization frameworks often struggle with expandability. Many of these systems do not easily accommodate new techniques or updates from ongoing research, leaving users with outdated or incomplete options. Our approach is designed to be expandable, with a structure that allows for the integration of new techniques as they become available. This ensures that our collection remains relevant and up-to-date, which is essential for keeping up with advancements in data privacy and anonymization. It also allows the user to include custom-made techniques that come from their experience.

To finalise our approach we need to add a recommender system that further assists users in selecting the best-fitting anonymization approach. Such a system would help users select the optimal technique based on specific criteria, such as achieving the highest level of privacy protection or finding the best trade-off between data quality and privacy. Imple-

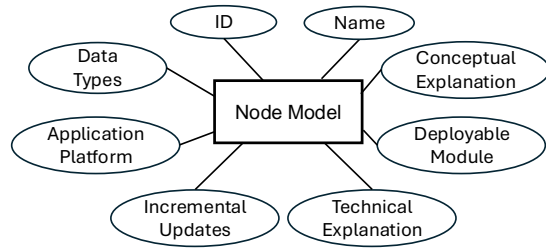


Figure 7: Core attributes of the model for the anonymization categories, techniques, and implementations.

menting this functionality would further enhance the practicality of our collection of anonymization techniques, making it easier for users to navigate the complex landscape of anonymization techniques.

Table 2: Model Attributes for Each Node.

Attributes	Description
ID	Unique identifier.
Name	Descriptive name of the node.
Conceptual Explanation	An overview building on the parent node’s explanation.
Data Types	Types of data the technique can handle (e.g., text, numerical, images).
Application Platform	Specifies whether the technique is for backend or on-board processing.
Incremental Updates	Indicates whether the technique supports updating datasets incrementally.
Implementation Details	Specific details about how the technique is implemented (e.g., algorithm used, coding languages).

6 TOWARDS PRACTICAL IMPLEMENTATION

In addition to the theoretical foundations discussed throughout this work, we have taken a significant step toward realizing a practical solution. We have developed an initial prototype of the anonymization toolbox and documented our efforts in a demo paper (Fieschi et al., 2025). This prototype serves as a proof of concept, showcasing the feasibility and potential of our approach.

Looking ahead, an essential component of the toolbox will be an integrated recommender sys-

tem. This system aims to support software developers by guiding them in selecting the most suitable anonymization techniques for their specific use cases. Furthermore, it will document the decision-making process, ensuring transparency and reproducibility in the selection of privacy-preserving methods. Our current research is now focused on finding possible solutions for such a recommender system to maximize its usability and effectiveness.

To evaluate the practicality and acceptance of the framework, we see the need for extensive testing with developers and real-world use cases. This evaluation will help us assess how well the toolbox aligns with the needs of its intended users and identify areas for further improvement. Our initial testing efforts will be conducted within the automotive domain, leveraging its complex and privacy-sensitive use cases as a foundation for iterative refinement of the collection of anonymization techniques.

7 CONCLUSIONS

This paper provides a comprehensive overview of anonymization techniques resulting from a systematic literature review and careful categorization of the available methods. By structuring these techniques into an organized framework, we offer users a valuable resource for making informed decisions about which anonymization approach best suits the data-collecting use case under development.

Our conceptual framework for anonymization techniques categorizes them into distinct clusters, making it easier to navigate through the various options and select the most appropriate method for specific use cases. This structured approach addresses the need for a clear and accessible overview of available anonymization techniques, supporting more effective decision-making in privacy protection. This lays a foundation for future research and implementations, enhancing the potential for anonymization by design, its application, and spread.

The presence of a recommender system would significantly improve our framework to guide users in selecting the optimal technique based on their specific requirements. Developing such a system represents a key area for future research, which could further enhance the practicality and effectiveness of our collection of anonymization techniques.

ACKNOWLEDGEMENTS

This work is based on the research project SofDCar (19S21002), funded by the German Federal Ministry for Economic Affairs and Climate Action.

REFERENCES

- Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G. S., Davis, A., Dean, J., Devin, M., Ghemawat, S., Goodfellow, I., Harp, A., Irving, G., Isard, M., Jia, Y., Jozefowicz, R., Kaiser, L., Kudlur, M., Levenberg, J., Mané, D., Monga, R., Moore, S., Murray, D., Olah, C., Schuster, M., Shlens, J., Steiner, B., Sutskever, I., Talwar, K., Tucker, P., Vanhoucke, V., Vasudevan, V., Viégas, F., Vinyals, O., Warden, P., Wattenberg, M., Wicke, M., Yu, Y., and Zheng, X. (2015). TensorFlow: Large-scale machine learning on heterogeneous systems. Software available from tensorflow.org.
- Abay, N. C., Zhou, Y., Kantarcioglu, M., Thuraisingham, B., and Sweeney, L. (2019). Privacy Preserving Synthetic Data Release Using Deep Learning. In Berlingerio, M., Bonchi, F., Gärtner, T., Hurley, N., and Ifrim, G., editors, *Machine Learning and Knowledge Discovery in Databases*, Lecture Notes in Computer Science, pages 510–526, Cham. Springer International Publishing.
- Abbasi, A. and Mohammadi, B. (2022). A clustering-based anonymization approach for privacy-preserving in the healthcare cloud. *Concurrency and Computation: Practice and Experience*, 34(1).
- Al-Kharji, S., Tian, Y., and Al-Rodhaan, M. (2018). A Novel (K, X)-isomorphism Method for Protecting Privacy in Weighted social Network.
- Aleroud, A., Shariah, M., and Malkawi, R. (2022). Privacy Preserving Human Activity Recognition Using Microaggregated Generative Deep Learning. In *2022 IEEE International Conference on Cyber Security and Resilience (CSR)*, pages 357–363.
- Aljably, R. (2021). *Privacy Preserving Data Sharing in Online Social Networks*, volume 1415 of *Communications in Computer and Information Science*. Pages: 152.
- Alnemari, A., Arodi, S., Sosa, V., Pandey, S., Romanowski, C., Raj, R., and Mishra, S. (2018). *Protecting infrastructure data via enhanced access control, blockchain and differential privacy*, volume 542 of *IFIP Advances in Information and Communication Technology*. Pages: 125.
- Aminifar, A., Rabbi, F., Pun, V., and Lamo, Y. (2021). Diversity-Aware Anonymization for Structured Health Data. pages 2148–2154.
- An, S., Li, Y., Wang, T., and Jin, Y. (2018). Contact Graph Based Anonymization for Geosocial Network Datasets. pages 132–137.
- Attaullah, H., Anjum, A., Kanwal, T., Malik, S., Asheralieva, A., Malik, H., Zoha, A., Arshad, K., and Imran,

- M. (2021). F-classify: Fuzzy rule based classification method for privacy preservation of multiple sensitive attributes. *Sensors*, 21(14).
- Chen, B.-C., Kifer, D., LeFevre, K., Machanavajjhala, A., et al. (2009). Privacy-preserving data publishing. *Foundations and Trends® in Databases*, 2(1–2):1–167.
- Domingo-Ferrer, J., Muralidhar, K., and Bras-Amoros, M. (2020). General Confidentiality and Utility Metrics for Privacy-Preserving Data Publishing Based on the Permutation Model. *IEEE Transactions on Dependable and Secure Computing*, pages 1–1.
- Dwork, C. (2008). Differential Privacy: A Survey of Results. In Agrawal, M., Du, D., Duan, Z., and Li, A., editors, *Theory and Applications of Models of Computation*, Lecture Notes in Computer Science, pages 1–19, Berlin, Heidelberg, Springer.
- Dwork, C. and Roth, A. (2013). The Algorithmic Foundations of Differential Privacy. *Foundations and Trends® in Theoretical Computer Science*, 9(3–4):211–407.
- European Parliament and Council of the European Union (2016). Regulation (EU) 2016/679 of the European Parliament and of the Council.
- Eyupoglu, C., Aydin, M., Zaim, A., and Sertbas, A. (2018a). An efficient big data anonymization algorithm based on chaos and perturbation techniques. *Entropy*, 20(5).
- Eyupoglu, C., Aydin, M., Zaim, A., and Sertbas, A. (2018b). An efficient big data anonymization algorithm based on chaos and perturbation techniques. *Entropy*, 20(5).
- Fan, H. and Wang, Y. (2023). Range optimal dummy location selection based on query probability density. *2023 2nd International Conference on Cloud Computing, Big Data Application and Software Engineering (CBASE)*, pages 366–371.
- Fieschi, A., Hirmer, P., Agrawal, S., Christoph, S., and Mitschang, B. (2024). Hysaad – a hybrid selection approach for anonymization by design in the automotive domain. In *2024 25th IEEE International Conference on Mobile Data Management (MDM)*. IEEE.
- Fieschi, A., Hirmer, P., and Stach, C. (2025). Discovering suitable anonymization techniques: A privacy toolbox for data experts. Presented at the 21st Conference on Database Systems for Business, Technology and Web (BTW 2025), Bamberg, Germany, March 2025.
- Fieschi, A., Li, Y., Hirmer, P., Stach, C., and Mitschang, B. (2023). Privacy in connected vehicles: Perspectives of drivers and car manufacturers. In *Symposium and Summer School on Service-Oriented Computing*, pages 59–68. Springer.
- Fung, B. C., Wang, K., Fu, A. W.-C., and Yu, P. S. (2010). Introduction to Privacy-Preserving Data Publishing. Chapman and Hall/CRC.
- Gaboardi, M., Hay, M., and Vadhan, S. (2020). A programming framework for opendp.
- Gao, T. and Li, F. (2019a). PHDP: Preserving Persistent Homology in Differentially Private Graph Publications. volume 2019-April, pages 2242–2250.
- Gao, T. and Li, F. (2019b). Privacy-Preserving Sketching for Online Social Network Data Publication. volume 2019-June.
- Gao, T. and Li, F. (2019c). Sharing Social Networks Using a Novel Differentially Private Graph Model.
- Hamm, J. (2017). Minimax filter: Learning to preserve privacy from inference attacks. *Journal of Machine Learning Research*, 18.
- Javed, I., Alharbi, F., Margaria, T., Crespi, N., and Qureshi, K. (2021). PETchain: A Blockchain-Based Privacy Enhancing Technology. *IEEE Access*, 9:41129–41143.
- Jin, F., Hua, W., Ruan, B., and Zhou, X. (2022). Frequency-based Randomization for Guaranteeing Differential Privacy in Spatial Trajectories. volume 2022-May, pages 1727–1739.
- Kalia, P., Bansal, D., and Sofat, S. (2021). Privacy Preservation in Cloud Computing Using Randomized Encoding. *Wireless Personal Communications*, 120(4):2847–2859.
- Khan, R., Tao, X., Anjum, A., Malik, S., Yu, S., Khan, A., Rehman, W., and Malik, H. (2022). (τ , m)-slicedBucket privacy model for sequential anonymization for improving privacy and utility. *Transactions on Emerging Telecommunications Technologies*, 33(6).
- Li, B., Zhu, H., and Xie, M. (2022). Releasing Differentially Private Trajectories with Optimized Data Utility. *Applied Sciences (Switzerland)*, 12(5).
- Li, N., Li, T., and Venkatasubramanian, S. (2007). t-Closeness: Privacy Beyond k-Anonymity and l-Diversity. In *2007 IEEE 23rd International Conference on Data Engineering*, pages 106–115. ISSN: 2375-026X.
- Li, X., Zhu, Y., and Wang, J. (2021). Highly Efficient Privacy Preserving Location-Based Services with Enhanced One-Round Blind Filter. *IEEE Transactions on Emerging Topics in Computing*, 9(4):1803–1814.
- Machanavajjhala, A., Kifer, D., Gehrke, J., and Venkatasubramanian, M. (2007). L-diversity: Privacy beyond k-anonymity. *ACM Transactions on Knowledge Discovery from Data*, 1(1):3–es.
- Mahdavifar, S., Deldar, F., and Mahdikhani, H. (2022). Personalized Privacy-Preserving Publication of Trajectory Data by Generalization and Distortion of Moving Points. *Journal of Network and Systems Management*, 30(1).
- Majeed, A. and Lee, S. (2021). Anonymization Techniques for Privacy Preserving Data Publishing: A Comprehensive Survey. *IEEE Access*, 9:8512–8545. Conference Name: IEEE Access.
- Mohamed, M., Ghanem, S., and Nagi, M. (2020). Privacy-preserving for distributed data streams: Towards l-diversity. *International Arab Journal of Information Technology*, 17(1):52–64.
- Moher, D., Shamseer, L., Clarke, M., Ghersi, D., Liberati, A., Petticrew, M., Shekelle, P., Stewart, L. A., and Group, P.-P. (2015). Preferred reporting items for systematic review and meta-analysis protocols (prisma-p) 2015 statement. *Systematic reviews*, 4:1–9.

- Morton, A. and Sasse, M. A. (2012). Privacy is a process, not a pet: a theory for effective privacy practice. In *Proceedings of the 2012 New Security Paradigms Workshop*, pages 87–104.
- Park, N., Mohammadi, M., Gorde, K., Jajodia, S., Park, H., and Kim, Y. (2018). Data synthesis based on generative adversarial networks. *Proceedings of the VLDB Endowment*, 11(10):1071–1083.
- Patki, N., Wedge, R., and Veeramachaneni, K. (2016). The synthetic data vault. In *2016 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*, pages 399–410.
- Pei, J., Xu, J., Wang, Z., Wang, W., and Wang, K. (2007). Maintaining k-anonymity against incremental updates. In *19th International Conference on Scientific and Statistical Database Management (SSDBM 2007)*, pages 5–5. IEEE.
- Piacentino, E. and Angulo, C. (2020a). *Anonymizing Personal Images Using Generative Adversarial Networks*, volume 12108 LNBI of *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. Pages: 405.
- Piacentino, E. and Angulo, C. (2020b). *Generating Fake Data Using GANs for Anonymizing Healthcare Data*, volume 12108 LNBI of *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. Pages: 417.
- Prasser, F. and Kohlmayer, F. (2015). Putting statistical disclosure control into practice: The arx data anonymization tool. Available at: <https://arx.deidentifier.org>.
- Puri, V., Kaur, P., and Sachdeva, S. (2022). (k, m, t)-anonymity: Enhanced privacy for transactional data. *Concurrency and Computation: Practice and Experience*, 34(18).
- Rong, H., Ma, T., Tang, M., and Cao, J. (2018). A novel subgraph K+-isomorphism method in social network based on graph similarity detection. *Soft Computing*, 22(8):2583–2601.
- Sei, Y. and Ohsuga, A. (2017). Location Anonymization with Considering Errors and Existence Probability. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 47(12):3207–3218.
- Shynu, P. G., Shayan., H. M., and Chowdhary, C. L. (2020). A fuzzy based data perturbation technique for privacy preserved data mining. *2020 International Conference on Emerging Trends in Information Technology and Engineering (ic-ETITE)*, pages 1–4.
- Stach, C. (2023). Data Is the New Oil—Sort of: A View on Why This Comparison Is Misleading and Its Implications for Modern Data Administration. *Future Internet*, 15(2):71:1–71:49.
- Sun, Y., Yuan, Y., Wang, G., and Cheng, Y. (2016). Splitting anonymization: a novel privacy-preserving approach of social network. *Knowledge and Information Systems*, 47(3):595–623.
- Sweeney, L. (2002). k-ANONYMITY: A MODEL FOR PROTECTING PRIVACY. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 10(05):557–570.
- Thouvenot, M., Curé, O., and Calvez, P. (2020). Knowledge graph anonymization using semantic anatomization. volume 2721, pages 129–133.
- Tsai, Y.-C., Wang, S.-L., Ting, I.-H., and Hong, T.-P. (2020). Flexible sensitive K-anonymization on transactions. *World Wide Web*, 23(4):2391–2406.
- Ullah, I. and Shah, M. (2016). A novel model for preserving Location Privacy in Internet of Things. pages 542–547.
- Walonoski, J., Kramer, M., Nichols, J., Quina, A., Moesel, C., Hall, D., Duffett, C., Dube, K., Gallagher, T., and McLachlan, S. (2017). Synthea: An approach, method, and software mechanism for generating synthetic patients and the synthetic electronic health care record. *Journal of the American Medical Informatics Association*, 25(3):230–238.
- Wang, Y., Wu, X., and Hu, D. (2016). Using Randomized Response for Differential Privacy Preserving Data Collection. In *EDBT/ICDT Workshops*.
- Ward, K., Lin, D., and Madria, S. (2017). MELT: Mapreduce-based efficient large-scale trajectory anonymization. volume Part F128636.
- Yamaç, M., Ahishali, M., Passalis, N., Raitoharju, J., Sankur, B., and Gabbouj, M. (2019). Reversible privacy preservation using multi-level encryption and compressive sensing. volume 2019-September.
- Yu, L., Liu, L., Pu, C., Gursoy, M. E., and Truex, S. (2019). Differentially Private Model Publishing for Deep Learning. In *2019 IEEE Symposium on Security and Privacy (SP)*, pages 332–349. arXiv:1904.02200 [cs].
- Zhao, H., Wan, J., and Chen, Z. (2016). A novel dummy-based knn query anonymization method in mobile services. *International Journal of Smart Home*, 10:137–154.
- Zhu, T., Li, G., Zhou, W., and Yu, P. S. (2017). Differentially Private Data Publishing and Analysis: A Survey. *IEEE Transactions on Knowledge and Data Engineering*, 29(8):1619–1638.