

Convergence of Differential Privacy and Shannon Entropy in Data Anonymization

Maria Virgínia Mendonça dos Santos

`virginia.msantos@ufape.edu.br`

Universidade Federal do Agreste de Pernambuco

Research Article

Keywords: Data Anonymization · Differential Privacy · Shannon Entropy · Systematic Literature Review · Privacy Metrics

Posted Date: October 22nd, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-11234407/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

Convergence of Differential Privacy and Shannon Entropy in Data Anonymization

A Systematic Literature Review

Maria Virgínia Mendonça dos Santos

Received: date / Accepted: date

Abstract The tension between data utility and individual privacy is a central challenge in computer science. While classical syntactic models such as *k-anonymity* have proven vulnerable to linkage and inference attacks, Differential Privacy (DP) emerged as the theoretical gold standard, offering robust mathematical guarantees regardless of the adversary's knowledge. The abstract complexity of DP, however, creates an intelligibility gap for decision-makers. This Systematic Literature Review, based on the analysis of 64 primary studies retrieved from three digital libraries, examines the state of the art in the evaluation of anonymization. The findings reveal hybridization rather than replacement: more than half of the studies (51.6%) combine mechanisms from distinct families, and DP has been appropriated less as an alternative to syntactic models than as an instrument for formally auditing them. Only 12.5% of the works propose mechanisms capable of translating formal guarantees for non-experts, and none of them addresses the data subject. Shannon Entropy, adopted by 31.3% of the studies, proves promising yet underused: it quantifies both the uncertainty of the adversary and the loss of information, which qualifies it as a semantic bridge between formal rigor and interpretability. The review concludes by proposing hybrid frameworks that integrate the rigor of DP with Conditional Entropy, sensitive to the auxiliary knowledge of the adversary, as a path toward auditable and transparent privacy systems.

Keywords Data Anonymization · Differential Privacy · Shannon Entropy · Systematic Literature Review · Privacy Metrics

1 Introduction

The release of databases plays a fundamental role in scientific progress and in the development of commercial strategies, yet it establishes an inherent tension with individual privacy. Historically, risk mitigation relied on the suppression of direct identifiers, a practice known as naive anonymization. The literature has since accumulated robust evidence of its ineffectiveness. Linkage attacks, such as the one seminally demonstrated by Sweeney [57], showed that the combination of quasi-identifiers (such as ZIP code, gender and date of birth) suffices to uniquely re-identify the vast majority of the population. The vulnerability worsens in high-dimensional settings, where the sparsity of the data acts as a fingerprint, as evidenced by the de-anonymization of the Netflix dataset by Narayanan and Shmatikov [38]. Such failures, theoretically underpinned by the reconstruction theorem of Dinur and Nissim [10], confirm Ohm's diagnosis [41] concerning the "broken promises" of simple anonymization, and establish the need for mathematically rigorous privacy guarantees.

In response to these vulnerabilities, the literature initially evolved towards models based on syntactic properties. The foremost among them was the notion of *k-anonymity* [48], which seeks to render a record indistinguishable from at least $k-1$ other records in the dataset. Although it became an industry standard, its limitations against homogeneity attacks motivated extensions such as *l-diversity* [32]. These syntactic approaches nevertheless share a fundamental flaw: they lack robust

Maria Virgínia Mendonça dos Santos (corresponding author)
Universidade Federal do Agreste de Pernambuco, Garanhuns,
Pernambuco, Brazil
ORCID: 0009-0001-0739-0123
E-mail: virginia.msantos@ufape.edu.br

security guarantees against adversaries equipped with background knowledge, which makes privacy contingent on what the attacker does not know.

The decisive break with this paradigm came with the formalization of Differential Privacy by Dwork et al. [12,13]. Unlike its predecessors, Differential Privacy abandons guarantees grounded in the structure of the data in favour of a mathematical guarantee over the analysis process, ensuring protection irrespective of the computational power or the prior knowledge of the adversary. This modern definition, a direct descendant of the advances in knowledge complexity and secure computation due to Goldwasser, Micali and Rackoff [20], established a new theoretical gold standard.

Despite the theoretical soundness of Differential Privacy, its practical implementation faces barriers of intelligibility and utility, creating a gap with respect to classical anonymization techniques which, although vulnerable, still predominate in certain settings. It is within this context of technological transition that the present work is situated. This Systematic Literature Review investigates how traditional models and their vulnerabilities have been reassessed in the light of formal guarantees, and examines whether Differential Privacy can act not only as a technical benchmark but as a tool for translating risk to decision-makers. Finally, the study examines the renewed interest in Shannon's Information Theory [50], assessing how Entropy may serve as a unifying metric for quantifying the conflict between privacy preservation and the analytical utility of data.

Against this background, the present work offers three main contributions. First, it provides a systematic mapping of the anonymization techniques adopted in the recent literature and of the residual vulnerabilities associated with each family of models, showing that the choice of a technique entails accepting a specific threat profile. Second, it empirically characterizes the dual role assumed by Differential Privacy, which has been appropriated less as a substitute for syntactic models than as an instrument for formally auditing them, and quantifies the scarcity of mechanisms that translate its guarantees for non-specialist audiences. Third, it systematizes the uses of Shannon Entropy as both a privacy and a utility metric, identifying it as a candidate semantic bridge between formal rigor and interpretability, and outlining Conditional Entropy as the formulation required to incorporate the auxiliary knowledge of the adversary.

The remainder of this paper is organized as follows. Section 2 describes the protocol of the systematic review, including the research questions, the search strategy, the selection criteria, the quality assessment and the data extraction and categorization procedures. Sec-

Table 1 Research questions.

ID	Research Question
RQ1	Which classical anonymization algorithms or techniques exist, how do they relate to privacy and data usability metrics, and which attacks or vulnerabilities affect these models?
RQ2	How can Differential Privacy be used as a formal evaluation of classical anonymization models, and how can it be translated into guarantees that are comprehensible to decision-makers and to the general public?
RQ3	How can Shannon entropy be used as a privacy or data usability metric in classical anonymization algorithms or techniques?

tion 3 presents and discusses the results obtained for each research question. Section 4 examines the threats to the validity of the study. Finally, Section 5 summarizes the conclusions and points to directions for future work.

2 Methodology

This Systematic Literature Review was conducted in accordance with the methodological guidelines established by Kitchenham and Charters [27]. The research protocol was structured with the aim of identifying, appraising and synthesizing the scientific evidence available on anonymization techniques, on the application of differential privacy and on the use of entropy as a validation metric. The entire process, from the registration of the protocol to the extraction of the data, was managed on the Parsifal platform, a tool supporting the conduct of systematic reviews, which made it possible to retain a complete trace of the screening decisions and to produce an auditable export of the final set of studies.

2.1 Research Questions

To guide the investigation, the main objective was decomposed into three specific Research Questions (RQs), presented in Table 1. These questions shaped the extraction strategy and the categorization of the results.

2.2 Search Strategy

Primary studies were retrieved through automated searches in three digital libraries, selected for their relevance and coverage in the area of computational security and privacy: ACM Digital Library, IEEE Xplore and Web of Science.

Table 2 Search strings applied per research question.

RQ	Search String
RQ1	("anonymization algorithms" OR "anonymization techniques") AND ("privacy metrics" OR "data usability metrics" OR "attacks on anonymization algorithms" OR "attacks" OR "vulnerabilities")
RQ2	("anonymization algorithms" OR "anonymization techniques") AND ("differential privacy" OR "formal evaluation" OR "understandable guarantees" OR "comprehensible guarantees" OR "decision-makers")
RQ3	("anonymization algorithms" OR "anonymization techniques") AND ("shannon's entropy" OR "shannon" OR "claude shannon" OR "information theory")

The search strategy followed a partitioned approach: specific search strings were constructed for each Research Question, using Boolean operators (AND, OR) to combine the key terms. Table 2 details the exact query strings applied.

The searches were executed in two sessions: IEEE Xplore and Web of Science were queried on 25 October 2025, and the ACM Digital Library on 29 October 2025. No temporal restriction was applied in the formulation of the queries; the range observed in the results (2006–2025) followed from the distribution of the retrieved studies themselves. The execution returned 1,280 records in total, distributed as follows: 904 from the ACM Digital Library (70.6%), 197 from Web of Science (15.4%) and 179 from IEEE Xplore (14.0%).

The records were exported in BibTeX format and imported into the Parsifal platform, which performed automatic duplicate detection by title comparison, flagging 418 redundant records. A manual check was additionally applied to the final set of selected studies, with prior normalization of the titles (removal of punctuation, standardization of typographic apostrophes and of spacing). This procedure identified two residual duplicates that the automatic detection had not captured, arising from character-encoding discrepancies between records exported by different databases. Both were removed, so that the final set corresponds to genuinely distinct studies.

2.3 Selection Criteria

To ensure the technical quality and the thematic adherence of the retrieved works, inclusion (Table 3) and exclusion (Table 4) criteria were established. The selection process took place in successive filtering stages: initial screening by title and abstract, reading of the

Table 3 Inclusion criteria.

ID	Inclusion Criterion
I1	Complete primary studies (full papers).
I2	Studies bearing a direct relationship to the focus of this work.
I3	Electronic availability of the full text.
I4	Publications written in English.
I5	Publications appearing in journals, symposia or conferences.

Table 4 Exclusion criteria.

ID	Exclusion Criterion
E1	Extended abstracts, short papers or incomplete works.
E2	Redundant study.
E3	Duplicate studies.
E4	Studies bearing no relationship to the focus of this work.
E5	Secondary (systematic reviews, surveys) and tertiary studies.
E6	Strictly application- or engineering-oriented focus without relevant theoretical contribution.
E7	Implementation described only superficially in the text.
E8	Book chapters or complete books.

introduction and conclusion and, finally, full-text reading. A study was excluded as soon as it satisfied any one of the exclusion criteria, without the need to advance to the subsequent stages.

It should be noted that screening, quality assessment and data extraction were carried out by a single researcher, so that measures of agreement between independent assessors, such as Cohen’s κ coefficient, could not be computed. To mitigate the risk of bias arising from this limitation, three procedures were adopted: the inclusion and exclusion criteria were registered in the protocol before screening began; all selection decisions were recorded on the Parsifal platform, so that the set of studies assessed at each stage is fully reconstructible; and studies whose eligibility remained uncertain at the title-and-abstract stage were promoted to the next stage rather than excluded, favouring sensitivity over precision in the early phases. The implications of this limitation are discussed in Section 4.

2.4 Quality Assessment

In addition to the eligibility criteria, a Quality Assessment checklist was applied in order to weigh the methodological soundness of the selected studies. Each item in Table 5 was scored according to a ternary scale: Yes (1.0), Partially (0.5) or No (0.0).

Table 5 Quality assessment criteria.

ID	Assessment Criterion
Q1	Is the study based on research?
Q2	Do the authors state the aim of the study clearly?
Q3	Is the proposed approach clearly described?
Q4	Is the research context clearly described?
Q5	Is the research context described at an adequate level of detail?
Q6	Is there a discussion of the results obtained?
Q7	Are the limitations of the study clearly described?
Q8	Is there a clear presentation of the open problems in the field?
Q9	Is there sufficient information for the study to be replicated?
Q10	Is the study supported by a tool or prototype?
Q11	Is there a clear justification for carrying out the study?

Table 6 Data extraction form.

RQ	Extracted Field
RQ1	Main anonymization technique or algorithm addressed.
RQ1	Metrics used to assess the <i>utility</i> of the data.
RQ1	Metrics used to assess the <i>privacy</i> of the data.
RQ1	Specific attack or vulnerability mentioned or mitigated.
RQ2	Relationship of the study to Differential Privacy.
RQ2	Proposal of mechanisms for translating privacy guarantees for non-experts or decision-makers.
RQ3	Manner in which Shannon Entropy, or metrics derived from it, is used.

The score of each study corresponds to the sum of the eleven items, and therefore ranges between 0.0 and 11.0. A minimum score of 7.0 was established as the cut-off point, corresponding to approximately 64% of the maximum score: studies that did not reach this threshold were excluded from the final set on grounds of insufficient methodological soundness, even when they had satisfied the criteria of thematic eligibility. Among the included studies, the observed scores range from 7.5 to 11.0, with a mean of 9.74 and a median of 10.0.

2.5 Data Extraction

For each study approved in the preceding stages, a structured extraction form comprising seven fields was applied, designed so that each field addressed a specific dimension of the research questions, as presented in Table 6. The form was implemented on the Parsifal platform and completed on the basis of the full-text reading of each study.

The fields were completed in free-text form, preserving the terminology employed by each primary study. This choice avoided framing the evidence prematurely within predefined categories, but required a subsequent coding stage, described in Section 2.6.

2.6 Categorization of Techniques

Synthesizing the extracted data required organizing the textual descriptions into comparable categories. An inductive approach was adopted: the categories were not defined a priori, but emerged from successive readings of the extracted fields, being refined and consolidated as new studies were incorporated, until the set stabilized. This procedure was adopted on account of the terminological heterogeneity observed across the primary studies, which frequently describe equivalent mechanisms under distinct designations.

Two characteristics of this coding must be made explicit, since they condition the reading of the quantitative results presented in Section 3. First, the assignment is *multi-label*: a single study may be associated with more than one category, since a substantial share of the works analysed propose hybrid approaches, combining, for instance, syntactic models with differential privacy mechanisms, or applying generalization techniques to graph-structured data. Consequently, the sum of the absolute frequencies of the categories exceeds the number of primary studies, and the proportions reported must be interpreted as the fraction of studies that address each category, and not as mutually exclusive parts of a whole. Second, for the comparative analysis of privacy and utility metrics, the emergent categories were consolidated into a smaller number of groups, as detailed in Section 3.4, so as to avoid fragmenting the sample into subsets without critical mass for comparison.

3 Results and Discussion

3.1 Study Selection

Executing the search strings returned 1,280 records. Duplicate removal eliminated 420 records (418 identified automatically by the platform and 2 detected in the manual check described in Section 2.2), leaving 860 studies for screening. Reading titles and abstracts excluded 592 records, yielding 268 studies considered potentially relevant. The subsequent stage, based on reading the introduction and the conclusion, excluded 192 studies, and full-text reading excluded a further 9, for a total of 67 studies submitted to quality assessment.

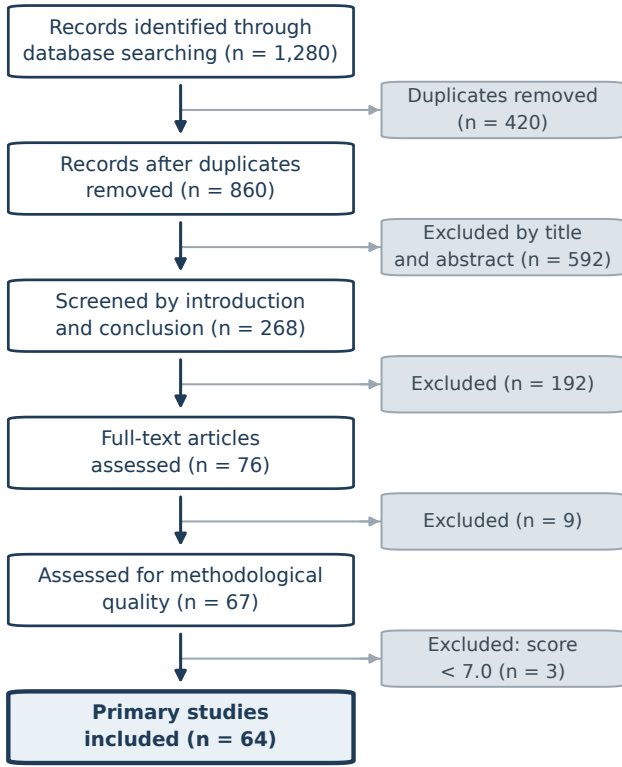


Fig. 1 PRISMA flow diagram detailing the study selection process

Of these, 3 did not reach the minimum score of 7.0 established in Section 2.4 and were discarded on grounds of insufficient methodological soundness. The final set therefore comprises 64 primary studies. The quantitative detail of each filtering stage is presented in the PRISMA flow diagram (Fig. 1).

3.2 Overview of the Studies

The longitudinal analysis (Fig. 2) shows that scientific output on anonymization has behaved in an oscillating manner over the two decades covered (2006–2025). The period of greatest productivity falls between 2011 and 2016, which accounts for 31 of the 64 studies (48.4%), with peaks in 2011 and 2015 (seven studies each), suggesting the phase in which the classical techniques were initially consolidated. From 2019 onward a sustained reduction in annual volume is observed, averaging fewer than three studies per year. It should be noted that 2018 shows a localized rise (six studies), coinciding with the entry into force of the General Data Protection Regulation (GDPR) in the European Union. This value, however, does not exceed those observed in 2011 and 2015, so that the data of this review do not support

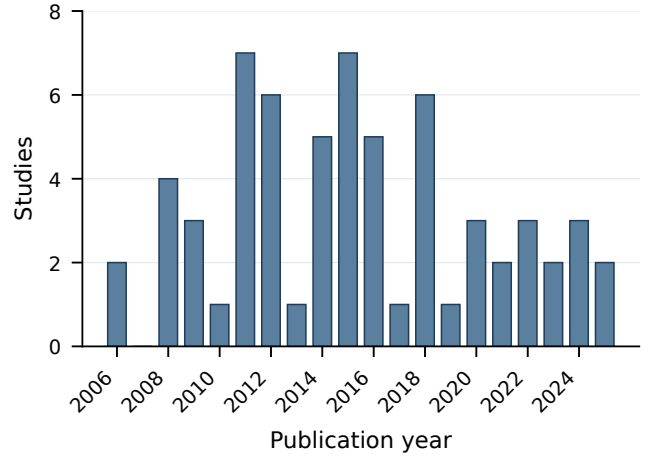


Fig. 2 Temporal distribution of the publications (2006–2025)

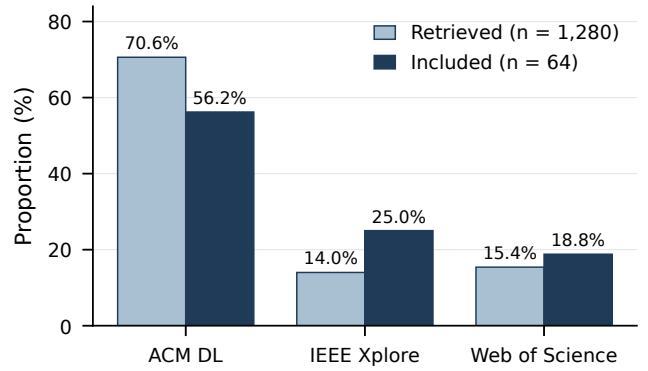


Fig. 3 Distribution by database: retrieved records ($n = 1,280$) and included primary studies ($n = 64$)

attributing a resurgence of the field to the regulatory change.

The distribution of the primary studies by source database (Fig. 3) differs appreciably from the distribution of the retrieved records. At the identification stage, the ACM Digital Library accounted for 70.6% of the 1,280 records (904), followed by Web of Science (197; 15.4%) and IEEE Xplore (179; 14.0%). After applying the eligibility criteria, however, the concentration falls substantially: among the 64 included studies, the ACM Digital Library accounts for 36 works (56.2%), IEEE Xplore for 16 (25.0%) and Web of Science for 12 (18.8%). This difference suggests that the greater coverage of the ACM in the initial retrieval reflects above all the sensitivity of the search strings to that database, rather than an equivalent concentration of the literature genuinely pertinent to the scope of this review.

The conferences indexed by the ACM (CCS, SIGMOD, KDD) have historically been the principal venues for disseminating security and privacy algorithms, which helps to explain their majority presence. Nevertheless,

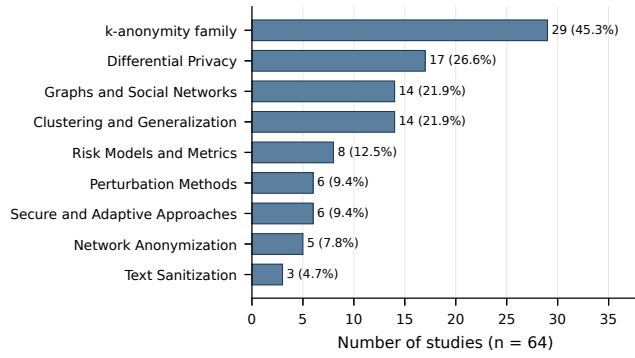


Fig. 4 Distribution of the primary studies by category of anonymization technique. The categories are not mutually exclusive: the proportions indicate the fraction of the 64 studies that addresses each category and therefore do not sum to 100%

the balance observed in the final set reinforces the need for caution in generalizing the results to interdisciplinary contexts, since the three databases privilege technical and computational output over legal or social approaches.

3.3 Models, Techniques and Vulnerabilities (RQ1)

The first Research Question sought to map the anonymization techniques that predominate in the contemporary literature, as well as to identify the intrinsic vulnerabilities that drive the development of new approaches. The analysis reveals a scenario of technological transition, characterized by the coexistence of mature classical syntactic models and emerging formal privacy guarantees.

The distribution of the techniques identified (Fig. 4) shows the predominance of the *k-anonymity* family, addressed by 29 of the 64 studies analyzed (45.3%). This prevalence may be attributed to the conceptual maturity of the model and to its low barrier to implementation in relational database management systems. Differential Privacy occupies second place, present in 17 studies (26.6%), followed by approaches directed at Graphs and Social Networks (14; 21.9%) and by *Clustering* and Generalization methods (14; 21.9%). The spectrum is completed by Risk Models and Metrics (8; 12.5%), Perturbation Methods (6; 9.4%), Secure and Adaptive Approaches (6; 9.4%), the anonymization of communication Networks (5; 7.8%) and Text Sanitization (3; 4.7%).

As stated in Section 2.6, this distribution is multi-label. Of the 102 assignments made over the 64 studies (a mean of 1.59 categories per study), 33 works (51.6%) were associated with two or more categories. This finding is in itself a relevant result: more than half of the output analyzed does not fall within a single paradigm,

but combines mechanisms from distinct families. Representative cases include the application of random sampling as a step preceding *k-anonymization* in order to obtain differential guarantees [29], the construction of differentially private decision trees from *k-anonymous* sets [19], and the combination of differential privacy with homomorphic encryption for genomic data [44]. The reading according to which there would be a linear replacement of syntactic models by Differential Privacy is therefore inadequate: what is observed is hybridization.

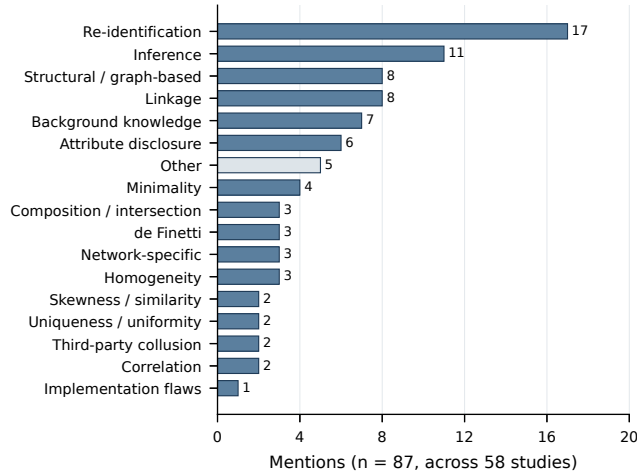
Table 7 establishes the link between the techniques employed and their residual vulnerabilities. The qualitative analysis indicates that, although the *k-anonymity* family is the most widespread, it remains susceptible to a broad vector of attacks, notably re-identification through linkage and the exploitation of background knowledge. This suggests that a significant share of the current literature is still devoted to mitigating the weaknesses of a syntactic model, seeking to optimize the trade-off between utility and privacy within known architectural limits. Differential Privacy models, by contrast, although they rigorously constrain linkage attacks, face challenges of another nature, such as membership inference, the practical difficulties of calibrating the privacy budget ϵ and, in at least one case, flaws in the implementation of the mechanisms themselves: floating-point vulnerabilities and defects in random number generation [9]. This distinction is fundamental, since it frames the choice of model not merely by the type of data, but by the capability profile of the adversary considered.

Coding the threats declared by the primary studies (Fig. 5) yielded 87 mentions distributed across 58 works; six studies do not make explicit the attack model they intend to mitigate. Re-identification remains the central risk, mentioned by 17 studies, but adversarial strategies evidently go beyond simple de-anonymization. Notable among them are Inference attacks ($n = 11$), aimed at the probabilistic deduction of sensitive attributes; Linkage attacks ($n = 8$) and the exploitation of Background knowledge ($n = 7$), which expose the fragility of static data when cross-referenced with external sources; and Structural attacks on graphs ($n = 8$). The occurrence of Composition and *de Finetti* attacks ($n = 3$ each), although less frequent, marks out critical research frontiers, indicating that continuous publication and topological analysis demand dynamic defense mechanisms superior to traditional approaches.

In sum, the analysis concerning RQ1 delineates a field in motion, stretched between the broad operational adoption of classical syntactic techniques and the pursuit of more robust formal guarantees. The mapping

Table 7 Summary of anonymization techniques and their main vulnerabilities.

Technique / Model Category	Main Vulnerabilities / Attacks
<i>K-Anonymity</i> Family	Re-identification, linkage attacks, homogeneity attacks, background knowledge attacks, structural attacks.
Differential Privacy	Membership inference, composition attacks, de-anonymization, attribute inference, implementation flaws.
<i>Clustering</i> and Generalization	Re-identification through quasi-identifiers, minimality attacks, foreground knowledge attacks, uniformity attacks.
Graph / Social Network Anonymization	Structure-based de-anonymization, neighborhood attacks, inference attacks.
Risk Models and Metrics	<i>de Finetti</i> attacks, composition attacks, re-identifying and revealing adversaries.
Secure and Adaptive Approaches	Inference attacks on APIs, third-party collusion, information fusion.
Network Anonymization	Predecessor attacks, fingerprinting, injection and frequency attacks, host re-identification.
Perturbation Methods	Structural de-anonymization, attacks on community structure, consumer re-identification.
Text Sanitization	Whistleblower re-identification, correlation attacks, background knowledge inference.

**Fig. 5** Attacks and vulnerabilities mentioned in the primary studies ($n = 87$ mentions across 58 studies; a single study may declare more than one threat)

shows that the selection of an anonymization model is intrinsically tied to accepting a residual vulnerability profile, requiring careful assessment of the trade-off between data utility, computational cost and the threat model under consideration.

3.4 Privacy and Utility Metrics

For the comparative analysis of evaluation metrics, the nine emergent categories were consolidated into five groups, as anticipated in Section 2.6. The consolidation was grounded in the conceptual overlap of the evaluation methods: Risk Models and Metrics, Secure and Adaptive Approaches and Text Sanitization were incorporated into the residual Perturbation group, given the shared probabilistic nature of their metrics and the low sample representativeness of the last; and the anonymization of communication Networks was merged with Graph Anonymization, owing to the identity of the topological properties analyzed. The multi-label rule was retained, so that the counts in Table 8 are consistent with those in Fig. 4 and, for the same reason, sum to more than 64.

The picture reveals a methodological heterogeneity marked by asymmetry. The preponderance of the *k-anonymity* family ($n = 29$) over Differential Privacy ($n = 17$) shows that anonymization practice remains anchored in approaches that depend on attack simulations (re-identification rates, for instance) in order to validate syntactic compliance (k, l). By contrast, DP offers theoretical rigor (ϵ), but dissociates that guarantee from the measurement of utility, creating an evaluative vacuum. Crucially, the Graph and *Clustering* groups display the introduction of Shannon Entropy as a dual metric, capable of quantifying both the uncertainty of the adversary and the loss of information. The absence of a cross-cutting metric, compounded by the fact that the most widely used technique lacks native formal guarantees, corroborates the communicability gap discussed below.

3.5 Differential Privacy: Formal Auditing and the Communicability Deficit (RQ2)

The investigation of the second Research Question reveals a functional dichotomy in the application of Differential Privacy. Of the 64 primary studies, 25 (39.1%) establish some relationship with DP, whereas 31 do not mention it and 8 did not permit classification on the basis of the available text. Among the 25 that address it, practical integration is the predominant modality ($n = 11$), followed by critical contrast with syntactic

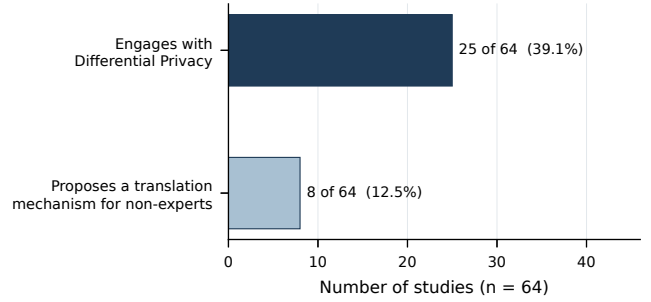
Table 8 Utility and privacy metrics by consolidated group of techniques. The groups are not mutually exclusive.

Group of Techniques	No.	Utility Metrics	Privacy Metrics
K-Anonymity Family (incl. <i>l-diversity</i> , <i>t-closeness</i> , <i>β-likeness</i>)	29	Information Loss (ILPG, AIL); classification/ <i>clustering</i> accuracy; structural similarity; number of operations.	Compliance (k, l, β); Privacy Gain (PG); Probability of Breach; re-identification accuracy.
Perturbation and Others (incl. risk models, secure approaches and text sanitization)	23	MILTON; compression ratio; recall/precision; time and scalability.	Algebraic proofs; resistance to third-party collusion; (p, N) risk metrics.
Graph and Network Anonymization	19	Structural similarity; edge metrics; Shannon Entropy (structural uncertainty).	Attack Success (Acc./FPR/AUC); Model Confidence ((k, l) -degree); Privacy ($1 - AC$); Shannon Entropy (de-anonymization).
Differential Privacy	17	Structural properties (<i>degree</i> , <i>path</i> , EVC); classification accuracy; overlaps/MAE.	ϵ -differential privacy; budget (ϵ, δ); global/local sensitivity; Anon Drop/PVP.
Clustering and Generalization	14	Information Loss (range query distortion); pattern preservation; Shannon Entropy (derived ILPG).	Privacy Gain (PG); <i>k-anon</i> / (g, l) -div; Shannon Entropy (privacy gain).

models ($n = 6$), use as a foundation or instrument of formal verification ($n = 4$), mention without further development ($n = 3$) and employment as a comparative benchmark ($n = 1$).

This distribution supports a reading that departs from the usual narrative. Although DP has historically been positioned as an alternative to syntactic models, the corpus shows that it has been progressively appropriated as a framework of formal verification for auditing the robustness of classical techniques. Studies such as those of Ganta et al. [17] and Xiao et al. [67] use the probabilistic bounds of DP, notably ϵ -indistinguishability, to quantify mathematically the structural vulnerabilities, such as composition and intersection attacks, that remained invisible under traditional deterministic metrics. In this context, DP operates less as a competitor than as a proof tool, offering the rigor needed to validate or refute the effectiveness of syntactic generalizations. The work of de Medeiros et al. [9] takes this strand to its limit, by mechanizing the formal verification of the very foundations of DP.

By contrast, and as illustrated in Fig. 6, the communicability of these guarantees remains a critical bottleneck. Only 8 of the 64 studies (12.5%) propose any mechanism for translating formal guarantees for non-specialist audiences. The mechanisms identified are heterogeneous and, for the most part, incidental to the main objective of the works: visual risk-utility frontiers that allow the data holder to choose the operating point [55]; the aggregation of multiple metrics into a single index through multi-criteria decision analysis [71]; discrete access levels associated with user profiles [42]; the explicit modeling of the probable number of adversaries as a translation of corporate risk [35]; the formulation

**Fig. 6** Discrepancy between the use of Differential Privacy as a technical auditing tool and the availability of translation mechanisms for non-experts

of guarantees in terms of relative confidence gain, which the authors argue is more intuitive than cumulative differences [7]; the customization of rigorous definitions by domain experts who are not privacy experts [26]; the recommendation to communicate the principles of DP in plain language in the user interface [31]; and the color-coded visualization of levels of concern for each passage of text, proposed explicitly as an interactive awareness tool [56].

It is further to be noted that none of the eight mechanisms addresses the data subject, the individual whose privacy is at stake. All of them are directed at the data holder, the analyst or the manager. A semantic gap at two levels thus becomes apparent: the literature has mastered the technique of protecting the data and has advanced in part in explaining that protection to those who operate it, but it has not developed instruments for communicating it to those whom it protects.

3.6 Adoption and Applicability of Shannon Entropy (RQ3)

The third Research Question examined the practical relevance of Shannon Entropy as a verification mechanism in anonymization processes. The analysis reveals a significant discrepancy between the theoretical grounding of the metric in Information Theory and its effective implementation in the recent literature.

From the quantitative perspective (Fig. 7a), the metric was explicitly adopted in 20 of the 64 studies (31.3%). This figure corroborates the predominance of empirical metrics, such as re-identification success rates, and of models based on rigid syntactic constraints, to the detriment of formal validations grounded in probabilistic uncertainty.

Qualitatively, a diversification of roles is observed (Fig. 7b). Among the 20 studies that employ it, 9 use it exclusively as a **privacy metric**, quantifying the uncertainty of the adversary with respect to the sensitive attributes, a scenario in which high entropy indicates robustness. A further 6 employ it exclusively as a **utility metric**, to measure Information Loss, where the entropic divergence between the original and the sanitized data is to be minimized. A third set, comprising 3 studies, uses the metric **in both functions simultaneously**, exploiting its duality within a single evaluative framework. The two remaining cases are singular and particularly informative: one incorporates entropy as an active component of the loss function [56] and the other subjects the sufficiency of the metric itself to critical scrutiny [4].

The trend represented by the first of these cases deserves emphasis, as it signals the transition of entropy from a passive metric of *a posteriori* assessment to an active component of optimization. Staufer et al. [56] integrate entropy into the loss function of sanitization models, guiding learning so as to maximize uncertainty over sensitive passages during training itself.

Despite this versatility, the sufficiency of entropy as a standalone metric remains under scrutiny. Arca and Hewett [4] warn that, without taking the prior knowledge of the adversary into account, the metric may fail to capture semantic risks. This limitation indicates that the robust adoption of Entropy requires the transition from the absolute uncertainty measure $H(X)$ to Conditional Entropy $H(X | K)$. In this formulation, the auxiliary knowledge K acts as a reducer of the probability space, permitting a semantic calibration in which privacy is assessed by the residual uncertainty of the adversary, and not merely by the statistical variability of the data.

4 Threats to Validity

This section discusses the limitations that condition the interpretation of the results presented, organized according to the dimensions of validity usually considered in systematic reviews.

4.1 Internal Validity

The principal limitation of this study lies in its having been conducted by a single researcher. All stages (screening, quality assessment, extraction and categorization) were carried out by one person, which makes it impossible to compute measures of agreement between independent assessors, such as Cohen's κ coefficient, and admits the influence of judgment bias in the decisions of inclusion and of coding.

Three procedures were adopted to mitigate this risk. First, the protocol (research questions, search strings, eligibility criteria and quality checklist) was registered in full before screening began, reducing the scope for *ad hoc* decisions. Second, the selection was conducted in four successive and progressively more costly stages (title, abstract, introduction and conclusion, and full text), so that no definitive exclusion occurred on the basis of minimal evidence; doubtful cases were promoted to the following stage, favoring sensitivity over precision in the early phases. Third, the protocol and the results were submitted to the supervision of an experienced researcher, who reviewed the methodological decisions. It should be noted, however, that supervision is not equivalent to independent double screening, and the limitation remains.

A second limitation concerns the audit trail. The formal recording, on the management platform, of the exclusion criterion applied to each study was carried out incompletely: of the discarded records, only a minor fraction had the corresponding criterion explicitly linked. The exclusion decision itself followed the criteria of Section 2.3 at every stage, but the published dataset does not make it possible to reconstruct, for each individual study, which criterion motivated its discard. Future work seeking to replicate or extend this review should record that association from the first screening stage onward.

It should finally be noted that the categorization of the techniques (Section 2.6) resulted from inductive coding performed by a single coder. Emergent categories are sensitive to the order of reading and to the interpretation of the coder, particularly when the descriptions given by the primary studies employ heterogeneous terminology for equivalent mechanisms. The complete table of assignments has been made available

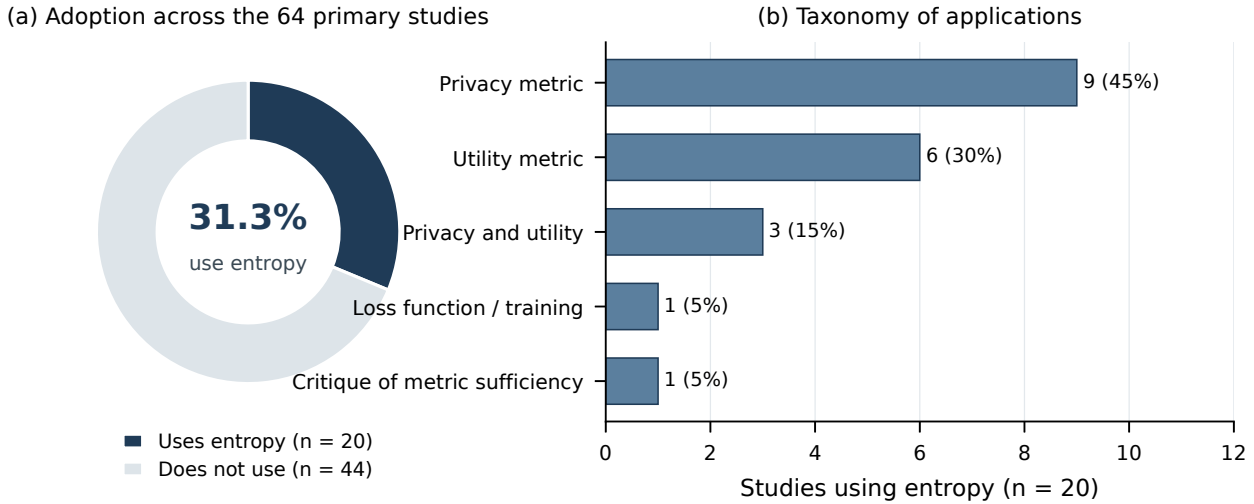


Fig. 7 Characterization of the use of Shannon Entropy: (a) proportion of adoption across the 64 primary studies and (b) taxonomy of the applications in the 20 studies that employ it

alongside the remaining data of the study, allowing each coding decision to be audited independently.

4.2 Construct Validity

Data extraction relied on free-text fields completed from the reading of the primary studies. This format preserved the original terminology, but left gaps: six studies do not make explicit the attack model they intend to mitigate, and eight did not permit classification as to their relationship with Differential Privacy. The frequencies reported in Figs. 5 and 6 must therefore be read as lower bounds rather than as exhaustive counts.

In addition, the identification of mechanisms for translating guarantees for non-experts (RQ2) depends on the primary study declaring that intention explicitly. It is plausible that some works present visualizations or parameterizations with a practical translation effect without framing them as such, which would underestimate the proportion of 12.5% reported in Section 3.5.

4.3 External Validity

The search was restricted to three digital libraries (ACM Digital Library, IEEE Xplore and Web of Science) and to publications written in English. Both restrictions privilege technical and computational output and may have excluded relevant contributions appearing in legal venues, in information systems journals not indexed in those databases, or in other languages. The absence of Scopus and SpringerLink from the set consulted is an acknowledged limitation.

This restriction is particularly sensitive for RQ2. The translation of formal guarantees for decision-makers is a problem that manifests itself at the intersection of computing, law and human-computer interaction, and the literature of the latter two areas is underrepresented in the scope adopted. The finding of a scarcity of translation mechanisms must therefore be interpreted as characteristic of the computational literature on anonymization, and not as a diagnosis of the field of privacy as a whole.

4.4 Conclusion Validity

The analyses presented are descriptive and exploratory in nature: no tests of statistical significance were applied, and the differences between category frequencies must not be interpreted as statistically demonstrated differences. It should further be stressed that the multi-label categorization, although more faithful to the hybrid nature of the works analyzed, precludes reading the proportions as parts of a whole, as cautioned in Section 3.3.

5 Conclusion and Future Work

This Systematic Literature Review analyzed 64 primary studies on anonymization techniques, revealing a technological transition marked by an asymmetry between theoretical advances and practical application. The results confirm that, despite the academic consensus on the structural limitations of syntactic models, the *k-anonymity* family retains a significant prevalence, present in 29 of the 64 works (45.3%). Such persistence indicates

that implementation feasibility continues to take precedence over the demand for formal robustness in many operational contexts.

The most telling finding, however, is not the prevalence of one paradigm over another, but their coexistence. More than half of the studies analyzed (33; 51.6%) combine mechanisms from distinct categories, articulating syntactic models, differential guarantees, generalization techniques and cryptographic methods in hybrid arrangements. The narrative of a linear replacement of the classical models by Differential Privacy does not hold empirically: what is observed is composition.

With regard to Differential Privacy, the review identified a functional dichotomy. Present in 25 of the 64 studies (39.1%), DP is employed predominantly as an instrument of mathematical verification for specialists, auditing the robustness of classical techniques rather than replacing them. By contrast, only 8 studies (12.5%) propose mechanisms capable of translating abstract parameters such as ϵ into interpretable risk indicators, and none of them addresses the data subject: all are directed at the data holder, the analyst or the manager. The scarcity of translation methods thus constitutes a significant barrier to the broad adoption of formal guarantees outside the academic sphere, and a twofold gap, since it does not even reach the individual whom the protection is meant to benefit.

As to the role of Information Theory, the investigation of Shannon Entropy revealed its potential as a unifying metric, albeit an underused one: adopted by 20 of the 64 studies (31.3%), it serves both to measure the uncertainty of the adversary and the degradation of the signal, and in three cases performs both functions simultaneously within the same framework. It is concluded that progress in the field depends on the development of hybrid evaluation frameworks that integrate the auditable robustness of Differential Privacy with the interpretability of entropic metrics.

5.1 Future Work

On the basis of the gaps identified and of the opportunity for integrating formalism and interpretability, the following directions are suggested:

1. **Development of a hybrid (DP-entropic) framework:** to formalize an evaluation framework that uses Differential Privacy to guarantee the upper bound on risk and Shannon Entropy to translate that risk into metrics of information loss and residual uncertainty, thereby facilitating the calibration of parameters by non-experts.
2. **Active incorporation of entropy into optimization:** to investigate the use of entropy not merely as a metric of *post-hoc* validation, but as a component of the loss function in sanitization algorithms, generalizing to structured data the approach observed in text sanitization [56].
3. **Semantic entropy sensitive to auxiliary knowledge:** to develop metrics based on Conditional Entropy $H(X | K)$ capable of weighing statistical uncertainty against the prior knowledge of the adversary, thereby overcoming the limitation pointed out by Arca and Hewett [4].
4. **Empirical validation with decision-makers:** to conduct studies with data managers and Data Protection Officers in order to assess whether entropy-based metrics afford a superior understanding of re-identification risks compared with the raw presentation of the parameters ϵ or k .
5. **Communication directed at the data subject:** to investigate translation mechanisms aimed at the protected individual, a gap for which this review identified no proposal whatsoever in the corpus analyzed.

Declarations

Funding. The author received no support from any organization for the submitted work.

Competing interests. The author declares no relevant financial or non-financial interests to disclose.

Data availability. The data supporting the findings of this study, namely the search strings, the complete set of screening records, the list of the 64 included primary studies, the raw data extraction form, the technique and entropy classifications and the quality assessment scores, are openly available in Zenodo at <https://doi.org/10.5281/zenodo.22926474>. No individual or sensitive data were collected.

Author contributions. The conception of the study, the definition of the review protocol, the execution of the searches, the screening and extraction of the data, the analysis of the results and the writing and critical revision of the manuscript were carried out by Maria Virgínia Mendonça dos Santos, who read and approved the final version.

Ethics approval. Not applicable. The research involved neither human participants nor animals.

References

1. Ahmadinejad, S.H., Fong, P.W.L., Safavi-Naini, R.: Privacy and Utility of Inference Control Mechanisms for Social Computing Applications. In: 11th ACM on Asia Conference on Computer and Communications Security, pp. 829–840. ACM, Xi'an China (2016). <https://doi.org/10.1145/2897845.2897878>
2. Aleroud, A., Yang, F., Pallaprolu, S.C., Chen, Z., Karabatis, G.: Anonymization of Network Traces Data through Condensation-based Differential Privacy. *Digit. Threat. Res. Pract.* **2**(4), 1–23 (2021). <https://doi.org/10.1145/3425401>
3. Anandan, B., Clifton, C.: Significance of Term Relationships on Anonymization. In: IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology, pp. 253–256. IEEE, Lyon, France (2011). <https://doi.org/10.1109/WI-IAT.2011.240>
4. Arca, S., Hewett, R.: Is Entropy enough for measuring Privacy? In: International Conference on Computational Science and Computational Intelligence (CSCI), pp. 1335–1340. IEEE, Las Vegas, NV, USA (2020). <https://doi.org/10.1109/CSCI51800.2020.00249>
5. Atzori, M., Bonchi, F., Giannotti, F., Pedreschi, D.: Towards low-perturbation anonymity preserving pattern discovery. In: 2006 ACM symposium on Applied computing, pp. 588–592. ACM, Dijon France (2006). <https://doi.org/10.1145/1141277.1141412>
6. Burkhart, M., Brauckhoff, D., May, M., Boschi, E.: The risk-utility tradeoff for IP address truncation. In: 1st ACM workshop on Network data anonymization, pp. 23–30. ACM, Alexandria Virginia USA (2008). <https://doi.org/10.1145/1456441.1456452>
7. Cao, J., Karras, P.: Publishing microdata with a robust privacy guarantee. *Proc. VLDB Endow.* **5**(11), 1388–1399 (2012). <https://doi.org/10.14778/2350229.2350255>
8. Cormode, G., Srivastava, D., Li, N., Li, T.: Minimizing minimality and maximizing utility: analyzing method-based attacks on anonymized data. *Proc. VLDB Endow.* **3**(1-2), 1045–1056 (2010). <https://doi.org/10.14778/1920841.1920972>
9. De Medeiros, M., Naveed, M., Lepoint, T., Kahsai, T., Ravitch, T., Zetsche, S., Joshi, A., Tassarotti, J., Albarghouthi, A., Tristan, J.B.: Verified Foundations for Differential Privacy. *Proc. ACM Program. Lang.* **9**(PLDI), 1094–1118 (2025). <https://doi.org/10.1145/3729294>
10. Dinur, I., Nissim, K.: Revealing information while preserving privacy. In: 22nd ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems (PODS), pp. 202–210. ACM, San Diego California USA (2003). <https://doi.org/10.1145/773153.773173>
11. Doka, K., Xue, M., Tsoumakos, D., Karras, P.: k-Anonymization by Freeform Generalization. In: 10th ACM Symposium on Information, Computer and Communications Security, pp. 519–530. ACM, Singapore Republic of Singapore (2015). <https://doi.org/10.1145/2714576.2714590>
12. Dwork, C., McSherry, F., Nissim, K., Smith, A.: Calibrating noise to sensitivity in private data analysis. In: Theory of Cryptography Conference (TCC), pp. 265–284. Springer, New York (2006). https://doi.org/10.1007/11681878_14
13. Dwork, C., Roth, A.: The algorithmic foundations of differential privacy. *Found. Trends Theor. Comput. Sci.* **9**(3-4), 211–407 (2014). <https://doi.org/10.1561/04000000042>
14. Efros, P., Buchmann, E., Englhardt, A., Böhm, K.: How to quantify the impact of lossy transformations on change detection. In: 27th International Conference on Scientific and Statistical Database Management, pp. 1–11. ACM, La Jolla California (2015). <https://doi.org/10.1145/2791347.2791371>
15. Fang, J., Li, A., Jiang, Q., Li, S., Han, W.: A Structure-Based De-Anonymization Attack on Graph Data Using Weighted Neighbor Match. In: IEEE Fourth International Conference on Data Science in Cyberspace (DSC), pp. 480–486. IEEE, Hangzhou, China (2019). <https://doi.org/10.1109/DSC.2019.00079>
16. Ganta, S.R., Acharya, R.: Adaptive data anonymization against information fusion based privacy attacks on enterprise data. In: 2008 ACM symposium on Applied computing, pp. 1075–1076. ACM, Fortaleza, Ceara Brazil (2008). <https://doi.org/10.1145/1363686.1363936>
17. Ganta, S.R., Kasiviswanathan, S.P., Smith, A.: Composition attacks and auxiliary information in data privacy. In: 14th ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 265–273. ACM, Las Vegas Nevada USA (2008). <https://doi.org/10.1145/1401890.1401926>
18. Gao, T., Li, F.: Differential Private Social Network Publication and Persistent Homology Preservation. *IEEE Trans. Netw. Sci. Eng.* **8**(4), 3152–3166 (2021). <https://doi.org/10.1109/TNSE.2021.3106673>
19. Ghinita, G., Karras, P., Kalnis, P., Mamoulis, N.: A framework for efficient data anonymization under privacy and accuracy constraints. *ACM Trans. Database Syst.* **34**(2), 1–47 (2009). <https://doi.org/10.1145/1538909.1538911>
20. Goldwasser, S., Micali, S., Rackoff, C.: The knowledge complexity of interactive proof systems. In: 17th ACM Symposium on Theory of Computing (STOC), pp. 291–304. ACM, Providence Rhode Island USA (1985). <https://doi.org/10.1145/22145.22178>
21. Hong, Y., Vaidya, J., Lu, H., Wu, M.: Differentially private search log sanitization with optimal output utility. In: 15th International Conference on Extending Database Technology, pp. 50–61. ACM, Berlin Germany (2012). <https://doi.org/10.1145/2247596.2247604>
22. Hua, J., Tang, A., Fang, Y., Shen, Z., Zhong, S.: Privacy-Preserving Utility Verification of the Data Published by Non-Interactive Differentially Private Mechanisms. *IEEE Trans. Inf. Forensics Secur.* **11**(10), 2298–2311 (2016). <https://doi.org/10.1109/TIFS.2016.2532839>
23. Ji, S., Du, T., Hong, Z., Wang, T., Beyah, R.: Quantifying Graph Anonymity, Utility, and De-anonymity. In: IEEE INFOCOM Conference on Computer Communications, pp. 1736–1744. IEEE, Honolulu, HI (2018). <https://doi.org/10.1109/INFOCOM.2018.8486356>
24. Ji, S., Du, W., Srivatsa, M., Beyah, R.: Structural Data De-Anonymization: Theory and Practice. *IEEE/ACM Trans. Netw.* **24**(6), 3523–3536 (2016). <https://doi.org/10.1109/TNET.2016.2536479>
25. Khavkin, M., Last, M.: Preserving Differential Privacy and Utility of Non-stationary Data Streams. In: IEEE International Conference on Data Mining Workshops (ICDMW), pp. 29–34. IEEE, Singapore, Singapore (2018). <https://doi.org/10.1109/ICDMW.2018.00012>
26. Kifer, D., Machanavajjhala, A.: Pufferfish: A framework for mathematical privacy definitions. *ACM Trans. Database Syst.* **39**(1), 1–36 (2014). <https://doi.org/10.1145/2514689>
27. Kitchenham, B., Charters, S.: Guidelines for performing systematic literature reviews in software engineering.

- Technical Report EBSE 2007-001, Keele University and Durham University Joint Report (2007).
28. Kotra, A., Eldosouky, A., Sengupta, S.: Every Anonymization Begins with k : A Game-Theoretic Approach for Optimized k Selection in k -Anonymization. In: International Conference on Advances in Computing and Communication Engineering (ICACCE), pp. 1–6. IEEE, Las Vegas, NV, USA (2020). <https://doi.org/10.1109/ICACCE49060.2020.9154925>
 29. Li, N., Qardaji, W., Su, D.: On sampling, anonymization, and differential privacy or, k -anonymization meets differential privacy. In: 7th ACM Symposium on Information, Computer and Communications Security, pp. 32–33. ACM, Seoul Korea (2012). <https://doi.org/10.1145/2414456.2414474>
 30. Li, N., Zhang, N., Das, S.K.: Relationship Privacy Preservation in Publishing Online Social Networks. In: IEEE Third International Conference on Privacy, Security, Risk and Trust and IEEE Third International Conference on Social Computing, pp. 443–450. IEEE, Boston, MA, USA (2011). <https://doi.org/10.1109/PASSAT/SocialCom.2011.191>
 31. Liu, Q., Shakya, R., Khalil, M., Jovanovic, J.: Advancing privacy in learning analytics using differential privacy. In: 15th International Learning Analytics and Knowledge Conference, pp. 181–191. ACM, Dublin Ireland (2025). <https://doi.org/10.1145/3706468.3706493>
 32. Machanavajjhala, A., Kifer, D., Gehrke, J., Venkitasubramaniam, M.: l -diversity: privacy beyond k -anonymity. ACM Trans. Knowl. Discov. Data **1**(1), 3:1–3:52 (2007). <https://doi.org/10.1145/1217299.1217302>
 33. Majeed, A., Khan, S., Hwang, S.O.: Towards Optimization of Privacy-Utility Trade-Off Using Similarity and Diversity Based Clustering. IEEE Trans. Emerg. Top. Comput. **12**(1), 368–385 (2024). <https://doi.org/10.1109/TETC.2023.3258528>
 34. Medková, J.: Anonymization of geosocial network data by the (k, l) -degree method with location entropy edge selection. In: 15th International Conference on Availability, Reliability and Security, pp. 1–8. ACM, Virtual Event Ireland (2020). <https://doi.org/10.1145/3407023.3409184>
 35. Mimoto, T., Kiyomoto, S., Tanaka, K., Miyaji, A.: (p, N) -identifiability: Anonymity under Practical Adversaries. In: IEEE Trustcom/BigDataSE/ICSS, pp. 996–1003. IEEE, Sydney, Australia (2017). <https://doi.org/10.1109/Trustcom/BigDataSE/ICSS.2017.343>
 36. Mohammady, M., Wang, L., Hong, Y., Louafi, H., Pourzandi, M., Debbabi, M.: Preserving Both Privacy and Utility in Network Trace Anonymization. In: 2018 ACM SIGSAC Conference on Computer and Communications Security, pp. 459–474. ACM, Toronto Canada (2018). <https://doi.org/10.1145/3243734.3243809>
 37. Mohammed, N., Chen, R., Fung, B.C., Yu, P.S.: Differentially private data release for data mining. In: 17th ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 493–501. ACM, San Diego California USA (2011). <https://doi.org/10.1145/2020408.2020487>
 38. Narayanan, A., Shmatikov, V.: Robust de-anonymization of large sparse datasets. In: IEEE Symposium on Security and Privacy (SP), pp. 111–125. IEEE, Oakland, CA, USA (2008). <https://doi.org/10.1109/SP.2008.33>
 39. Nergiz, M.E., Tamersoy, A., Saygin, Y.: Instant anonymization. ACM Trans. Database Syst. **36**(1), 1–33 (2011). <https://doi.org/10.1145/1929934.1929936>
 40. Nilizadeh, S., Kapadia, A., Ahn, Y.Y.: Community-Enhanced De-anonymization of Online Social Networks. In: 2014 ACM SIGSAC Conference on Computer and Communications Security, pp. 537–548. ACM, Scottsdale Arizona USA (2014). <https://doi.org/10.1145/2660267.2660324>
 41. Ohm, P.: Broken promises of privacy: responding to the surprising failure of anonymization. UCLA Law Rev. **57**(6), 1701–1777 (2010).
 42. Palanisamy, B., Liu, L., Zhou, Y., Wang, Q.: Privacy-Preserving Publishing of Multilevel Utility-Controlled Graph Datasets. ACM Trans. Internet Technol. **18**(2), 1–21 (2018). <https://doi.org/10.1145/3125622>
 43. Panchenko, A.: On the Impact of Cross-Layer Information Leakage on Anonymity in Crowds. In: 11th ACM Symposium on QoS and Security for Wireless and Mobile Networks, pp. 35–42. ACM, Cancun Mexico (2015). <https://doi.org/10.1145/2815317.2815325>
 44. Raisaro, J.L., Choi, G., Pradervand, S., Colsonet, R., Jacquemont, N., Rosat, N., Mooser, V., Hubaux, J.P.: Protecting Privacy and Security of Genomic Data in i2b2 with Homomorphic Encryption and Differential Privacy. IEEE/ACM Trans. Comput. Biol. Bioinform. **15**(5), 1413–1426 (2018). <https://doi.org/10.1109/TCBB.2018.2854782>
 45. Ranjan, A., Ranjan, P.: Two-phase entropy based approach to big data anonymization. In: International Conference on Computing, Communication and Automation (ICCCA), pp. 76–81. IEEE, Greater Noida, India (2016). <https://doi.org/10.1109/CCAA.2016.7813693>
 46. Rodhe, I., Rohner, C., Ngai, E.C.H.: On location privacy and quality of information in participatory sensing. In: 8th ACM symposium on QoS and security for wireless and mobile networks, pp. 55–62. ACM, Paphos Cyprus (2012). <https://doi.org/10.1145/2387218.2387229>
 47. Sadhya, D., Chakraborty, B.: Quantifying the Effects of Anonymization Techniques Over Micro-Databases. IEEE Trans. Emerg. Top. Comput. **10**(4), 1979–1992 (2022). <https://doi.org/10.1109/TETC.2022.3141754>
 48. Samarati, P., Sweeney, L.: Protecting privacy when disclosing information: k -anonymity and its enforcement through generalization and suppression. Technical Report SRI-CSL-98-04, SRI International, Menlo Park (1998).
 49. Sei, Y., Ohsuga, A.: Randomized addition of sensitive attributes for l -diversity. In: 11th International Conference on Security and Cryptography (SECRYPT), pp. 1–11. SciTePress, Vienna, Austria (2014).
 50. Shannon, C.E.: A mathematical theory of communication. Bell Syst. Tech. J. **27**(3), 379–423 (1948). <https://doi.org/10.1002/j.1538-7305.1948.tb01338.x>
 51. Sherkat, R., Li, J., Mamoulis, N.: Efficient Time-Stamped Event Sequence Anonymization. ACM Trans. Web **8**(1), 1–53 (2013). <https://doi.org/10.1145/2532643>
 52. Silva, H., Basso, T., Moraes, R., Elia, D., Fiore, S.: A Re-Identification Risk-Based Anonymization Framework for Data Analytics Platforms. In: 14th European Dependable Computing Conference (EDCC), pp. 101–106. IEEE, Iasi (2018). <https://doi.org/10.1109/EDCC.2018.00026>
 53. Silva, M.J., Rijo, P., Francisco, A.: Evaluating the Impact of Anonymization on Large Interaction Network Datasets. In: First International Workshop on Privacy and Security of Big Data, pp. 3–10. ACM, Shanghai China (2014). <https://doi.org/10.1145/2663715.2669610>
 54. Soria-Comas, J., Domingo-Ferrer, J., Sanchez, D., Martinez, S.: t -Closeness through Microaggregation: Strict

- Privacy with Enhanced Utility Preservation. *IEEE Trans. Knowl. Data Eng.* **27**(11), 3098–3110 (2015). <https://doi.org/10.1109/TKDE.2015.2435777>
55. Srivastava, A., Geethakumari, G.: Determining privacy utility trade-off for Online Social Network data publishing. In: Annual IEEE India Conference (INDICON), pp. 1–6. IEEE, New Delhi, India (2015). <https://doi.org/10.1109/INDICON.2015.7443693>
 56. Staufer, D., Pallas, F., Berendt, B.: Silencing the Risk, Not the Whistle: A Semi-automated Text Sanitization Tool for Mitigating the Risk of Whistleblower Re-Identification. In: The 2024 ACM Conference on Fairness Accountability and Transparency, pp. 733–745. ACM, Rio de Janeiro Brazil (2024). <https://doi.org/10.1145/3630106.3658936>
 57. Sweeney, L.: k -anonymity: a model for protecting privacy. *Int. J. Uncertain. Fuzziness Knowl.-Based Syst.* **10**(5), 557–570 (2002). <https://doi.org/10.1142/S0218488502001648>
 58. Tachioka, Y.: Privacy preservation satisfying utility requirements based on multi-objective optimization. In: Joint 12th International Conference on Soft Computing and Intelligent Systems and 23rd International Symposium on Advanced Intelligent Systems (SCIS&ISIS), pp. 1–4. IEEE, Ise, Japan (2022). <https://doi.org/10.1109/SCISISIS55246.2022.10002081>
 59. Tassa, T., Gudes, E.: Secure distributed computation of anonymized views of shared databases. *ACM Trans. Database Syst.* **37**(2), 1–43 (2012). <https://doi.org/10.1145/2188349.2188353>
 60. Terrovitis, M., Mamoulis, N., Kalnis, P.: Privacy-preserving anonymization of set-valued data. *Proc. VLDB Endow.* **1**(1), 115–125 (2008). <https://doi.org/10.14778/1453856.1453874>
 61. Tsai, Y.C., Wang, S.L., Song, C.Y., Ting, I.H.: Privacy and Utility Effects of k -anonymity on Association Rule Hiding. In: The 3rd Multidisciplinary International Social Networks Conference on SocialInformatics 2016, Data Science 2016, pp. 1–6. ACM, Union NJ USA (2016). <https://doi.org/10.1145/2955129.2955169>
 62. Utaliyeva, A., Shin, J., Choi, Y.H.: Task-Specific Adaptive Differential Privacy Method for Structured Data. *Sensors* **23**(4), 1980 (2023). <https://doi.org/10.3390/s23041980>
 63. Vilalunga, A., Resende, J.S., Domingos, H.: TorKameleon: Improving Tor’s Censorship Resistance with K -anonymization and Media-based Covert Channels. In: IEEE 22nd International Conference on Trust, Security and Privacy in Computing and Communications (TrustCom), pp. 1490–1495. IEEE, Exeter, United Kingdom (2023). <https://doi.org/10.1109/TrustCom60117.2023.00203>
 64. Wang, S.L., Tsai, Y.C., Kao, H.Y., Hong, T.P.: K -anonymity on sensitive transaction items. In: IEEE International Conference on Granular Computing, pp. 723–727. IEEE, Kaohsiung, Taiwan (2011). <https://doi.org/10.1109/GRC.2011.6122687>
 65. Waseda, A., Nojima, R., Wang, L.: A Differentially Private (Random) Decision Tree without Noise from k -Anonymity. *Appl. Sci.* **14**(17), 7625 (2024). <https://doi.org/10.3390/app14177625>
 66. Wong, R.C.W., Fu, A.W.C., Wang, K., Yu, P.S., Pei, J.: Can the Utility of Anonymized Data be Used for Privacy Breaches?. *ACM Trans. Knowl. Discov. Data* **5**(3), 1–24 (2011). <https://doi.org/10.1145/1993077.1993080>
 67. Xiao, Q., Chen, R., Tan, K.L.: Differentially private network data release via structural inference. In: 20th ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 911–920. ACM, New York New York USA (2014). <https://doi.org/10.1145/2623330.2623642>
 68. Xu, J., Wang, W., Pei, J., Wang, X., Shi, B., Fu, A.W.C.: Utility-based anonymization for privacy preservation with less information loss. *SIGKDD Explor. Newsl.* **8**(2), 21–30 (2006). <https://doi.org/10.1145/1233321.1233324>
 69. Yakout, M., Atallah, M.J., Elmagarmid, A.: Efficient and Practical Approach for Private Record Linkage. *J. Data Inf. Qual.* **3**(3), 1–28 (2012). <https://doi.org/10.1145/2287714.2287715>
 70. Zakerzadeh, H., Aggarwal, C.C., Barker, K.: Privacy-preserving big data publishing. In: 27th International Conference on Scientific and Statistical Database Management, pp. 1–11. ACM, La Jolla California (2015). <https://doi.org/10.1145/2791347.2791380>
 71. Zhao, Y., Wagner, I.: Using Metrics Suites to Improve the Measurement of Privacy in Graphs. *IEEE Trans. Dependable Secure Comput.* **19**(1), 259–274 (2022). <https://doi.org/10.1109/TDSC.2020.2980271>
 72. Zhou, B., Han, Y., Pei, J., Jiang, B., Tao, Y., Jia, Y.: Continuous privacy preserving publishing of data streams. In: 12th International Conference on Extending Database Technology: Advances in Database Technology, pp. 648–659. ACM, Saint Petersburg Russia (2009). <https://doi.org/10.1145/1516360.1516435>
 73. Zhou, B., Pei, J.: The k -anonymity and l -diversity approaches for privacy preservation in social networks against neighborhood attacks. *Knowl. Inf. Syst.* **28**(1), 47–77 (2011). <https://doi.org/10.1007/s10115-010-0311-2>
 74. Zhou, Y., Shou, L., Shang, X., Chen, K.: Dissemination of anonymized streaming data. In: 9th ACM International Conference on Distributed Event-Based Systems, pp. 104–115. ACM, Oslo Norway (2015). <https://doi.org/10.1145/2675743.2771837>
 75. Zou, L., Chen, L., Özsu, M.T.: k -automorphism: a general framework for privacy preserving network publication. *Proc. VLDB Endow.* **2**(1), 946–957 (2009). <https://doi.org/10.14778/1687627.1687734>