Towards an Artifact to Assess Differential Privacy in Microdata Streams

Sean McElroy[1], Varghese Vaidyan[1] and Gurcan Comert[2]
[1]Dakota State University, Madison, USA
[2]North Carolina A&T State University, Greensboro, USA

me@seanmcelroy.com
Varghese.Vaidyan@dsu.edu
gcomert@ncat.edu

Abstract: As continued data breaches allow state-level threat actors to assemble expansive dossiers on populations to carry out information warfare objectives, protecting personal privacy in published data sets and internal data stores is increasingly essential to civilian and societal safety. At the same time, the explosion of high-resolution, high-accuracy microdata streams, such as timestamped geolocation coordinates collected simultaneously by hardware platforms, operating systems, and a multitude of on-device applications and sites establishes a layered, highly-correlated pattern of life that can uniquely identify individuals and allow for targeted information warfare actions. Differential privacy (DP) is an advanced but highly effective technique in protecting sensitive data streams. This robust approach preserves privacy in published data sets through additive statistical noise sampled from Gaussian or Laplacian probability distributions. Data sets that contain highly correlated event-based data require specialized techniques to preserve mathematical DP guarantees in microdata streams beyond "user-level" applications available in most off-the-shelf approaches. Because practitioners need more tools to assess the robustness of differentially private outputs in microdata streams, application errors may result in future reidentification and privacy loss for data subjects. This research yields an artifact that can reassociate events in microdata streams when insufficient naive approaches are used. It also serves as a tool for implementers to validate their approaches in highly correlated event data.

Keywords: Information warfare, Differential privacy, Microdata streams, Event-level privacy

1. Introduction

Differential privacy (DP) is an approach for applying perturbations with additive noise mechanisms to data sets to preserve an individual's data privacy in balance with the utility gained by publishing them for public inspection and analysis. DP is formally defined and provides a mathematical guarantee of privacy per the trade-off between individual privacy and publication utility set by a privacy loss budget parameter, ε. Often referred to as the "gold standard" for privacy preservation (Jarmin, 2019), this parameter has been the focus of much research as practitioners apply DP approaches across various domains, industries, and data types. When DP approaches are applied to data sets that include the exact individual many times with varying attributes, value shifts across different measurements can identify or provide for behavioural inferences that can be correlated to external data to identify individuals (Dwork et al, 2010; Xiao & Xiong, 2015). Literature provides varying methods to provide "event-level" DP to guard against this effect, ranging from how ε is determined to domain-specific additive noise implementation considerations. DP is a complex approach that is highly domain-specific and requires a thorough mathematical understanding to apply safely and preserve data privacy for individual-level DP, and even more so for event-level DP. For this reason, an artifact that assesses the robustness of a DP application over a microdata stream would be a beneficial tool to provide privacy assurance. This work introduces the Adversarial Event-Level Differential Privacy Assessor (AELDPA) tool that allows practitioners to validate ε and algorithmic choices against expected privacy protection outcomes.

2. Literature Review

Differential privacy has grown, matured, and evolved significantly in the nearly two decades since Dinur and Nissim (2003) revealed the 'Fundamental Law of Information Recovery', which asserts that one cannot guarantee privacy for a dataset without adding noise that perturbs the attributes of published elements. The advent of a mathematically sound method, which forms the basis of Cynthia Dwork's work on differential privacy, fundamentally changed technical data privacy and established the foundations of the field (Dwork, 2008; Dwork et al, 2010; Dwork & Roth, 2013). From these significant and fundamental contributions, the field has grown in many different research domains with varying approaches.

2.1 Two-Dimensional Approaches

When timestamps are collected within microdata streams, location data sets require "event-level differential privacy" (Dwork et al, 2010). The contribution of Xiao & Xiong (2015) regarding planar isotropic mechanism

(PIM) is that the classical notion of differential privacy must be made more specific to preserve location privacy using specialized techniques. Rather than requiring that a single event's membership be unknowable between two `neighbouring databases', a two-dimensional convex hull approach provides more stringent protection. For location-based, time-coded data, the event for a given subject must be indistinguishable from a δ-location set, defined as the probable locations for a user in their trajectory vector, rather than the entire database (Xiao & Xiong, 2015).

This contribution was novel and essential for many microdata stream manifestations, often encoded with absolute or relative timestamps. The approach "guarantees the true location is always protected in δ-location set at every timestamp" (Xiao & Xiong, 2015); however, it fundamentally requires that the data set publisher apply differential privacy to accurately define and implement the δ-location set membership for the correct definition of "probable" in the context of the microdata. For instance, pedometer data may define `probable' as the distance achievable by human movement. However, were this the sole definition of `probable', a jogger who embarks on a city bus that travels faster than a person could run would break this assumption. Regardless, this contribution provides a practical approach for a `continual observation' consideration first acknowledged by Dwork (Dwork et al, 2010; Dwork & Roth, 2013). The researchers also note that this method protects the data of a single event but does not protect the entire trajectory, which may be discernible from outside context or correlation even when using a δ-location set. Xiao & Xiong further developed a version of this publication in a journal article, which included a more focused look at Markov models (Xiao et al, 2017).

The introduction of IDF-OPT by Zhao et al (2021) is an essential development for privacy preservation for microdata streams. This work posits the challenges differential privacy faces when the underlying data set not only contains user trajectories that can be correlated with external data sets but also when the same subject is represented with multiple, including overlapping and repeated, trajectories that can allow for the establishment of behavioural patterns that allow for reidentification Zhao et al (2021). The approach and algorithms documented in this work are significant for research questions pertinent to microdata stream privacy, given that the model incorporates a correlation measurement and feedback mechanism to adjust additive noise parameters. While the method is specific to spatiotemporal trajectories, the general approach generalizes other highly correlated microdata data formats across differing dimensions outside spatiotemporal concerns. This quantitative, technical action research publication considers a tuple of time, latitude, and longitude, although the method is generally applicable across other dimensions of quasi-identifiers. The approach is algorithmic and treated as a Pareto optimization problem, which can be expressed in a format simple algorithms or even Boolean satisfiability testing tools for specialized use cases could address.

2.2 Domain-Specific Approaches

Zhen et al (2020) identified several mechanisms to extend differential privacy to itemsets to address the unique privacy leak cases while maintaining an improved balance of privacy budgets and utility at scale. Their work focuses on association rulemaking, a form of rule-based machine learning for identifying dependencies and relationships between variables (Agrawal, Imieliński, Swami, 1993). Similar in some respects to earlier research into frequent itemset mining, such as PrivBasis (Li et al, 2012), this form of machine learning frequently operates over microdata streams or event-based data, such as transactional records, to optimize sales, marketing, and operational efficiencies. Minimizing data for publication is beneficial to adhere to privacy best practices (Hoepman, 2020), whether the reduction of dimensionality improves the results of a specific transform, like differential privacy. It is important to consider that practical contributions should not be overfitted for a specific domain where generalized approaches with more applicability and reusability are possible.

2.3 Pattern Recurrence

Genomic data is not typically considered in the standard literature of microdata streams; however, it is a form of stream-based vector that is highly correlated. This work by Yilmaz et al (2022) explores not only the correlations present in genomic data but the domain space that allows for inference attacks on the application of differential privacy. Because genomic data is highly probable for sets of 64 codon sequences characteristics of amino acids and stop-signals, additive noise may not only be detectable when applied but could be filtered out using error correction techniques when insufficient privacy budgets are applied. The key contribution of this quantitative, experimental research is that noise may not be sufficient to apply across a vector of long, highly correlated, low-range values that have a high degree of pattern recurrence. Instead, domain knowledge about permissible sub-sequences in the vector may be needed to construct unique probability distributions to apply noise in a manner that domain awareness does not subjugate. While literature often considers

microdata streams in a few archetypal forms: click-streams, location pings or user trajectories, or biometric signals, vectors are an example of a microdata stream that is horizontally correlated for the entry in addition to potential repeated presence in a database. In addition to the specialized treatment to preserve privacy per the budget, this research also contributes an improvement on “randomized response”, often explicitly employed in local differential privacy to ensure plausible deniability in data that may be far more homogeneous than a full population data set.

3. Methodology

3.1 Technical Action Research

The development of this artifact used repeated design and empirical cycles as part of the technical action research (TAR) methodology. TAR is an interactive and iterative process that uses the outputs of an engineering cycle to help solve a client's engineering cycle (Wieringa, 2014). While the implementation of the artifact as part of a TAR client engineering cycle is beyond this scope, iterative development yielded cycles of generating microdata streams protected with ε-differential privacy (ε-DP) with increasing concurrency and realism in a synthesis model and cycles of developing an adversarial artifact that attempted to reassociate item-level events.

3.2 Data Synthesis

The research question for this effort was whether and to what extent an adversarial reidentification mechanism might be able to reidentify 3-tuple spatiotemporal microdata events, such as simulated geolocation data, containing a timestamp and X-Y coordinates when interspersed in a stream with multiple agents. While rich sources of real-world geographic microdata streams exist, this research tested a model of synthetically created data so the pre-treated and ε-DP treated views of microdata streams could be validated with complete knowledge to determine the performance of the adversarial artifact. The synthesis tool is written in .NET with C# and is hosted in a Godot project to provide visualizations of the synthesis (as well as adversarial reidentification). Synthesized data is output to a CSV file with columns for the agent ID, timestamp, X, and Y coordinates, as illustrated in Figure 1 and Figure 2. It generated coordinates for up to 20 concurrent agents moving over a 2D planar graph of nodes illustrated in Figure 3. Each agent was assigned a randomly generated list of must-visit nodes that started at a blue 'home' node, optionally went to a red node, to an orange node, optionally up to three green nodes, and traversed back to its blue home node. The path-finding algorithm used to construct the node order was A* between any two must-visit nodes, which provided for efficient pathing and could exhibit doubling-back behaviour across the entire trip (Hart, Nilsson, Raphael, 1968).

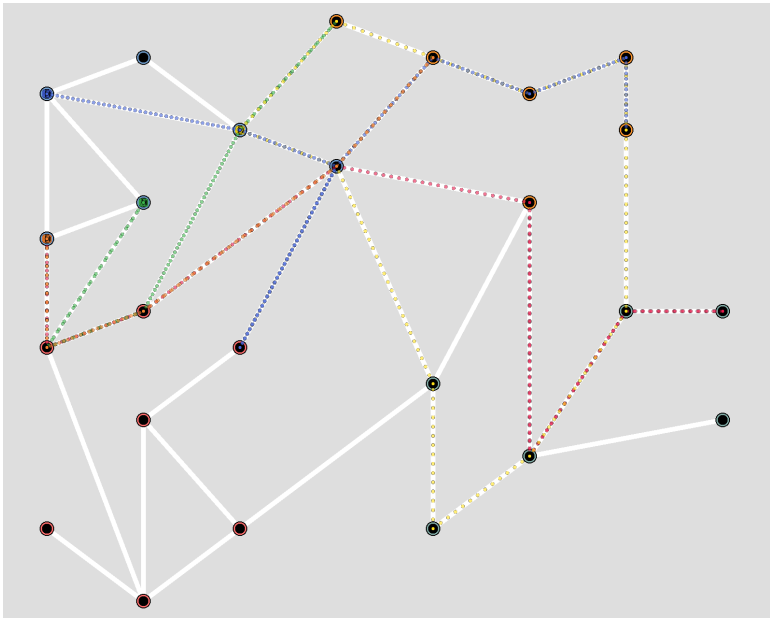

Figure 1: Model graph with 5 agent paths rendered

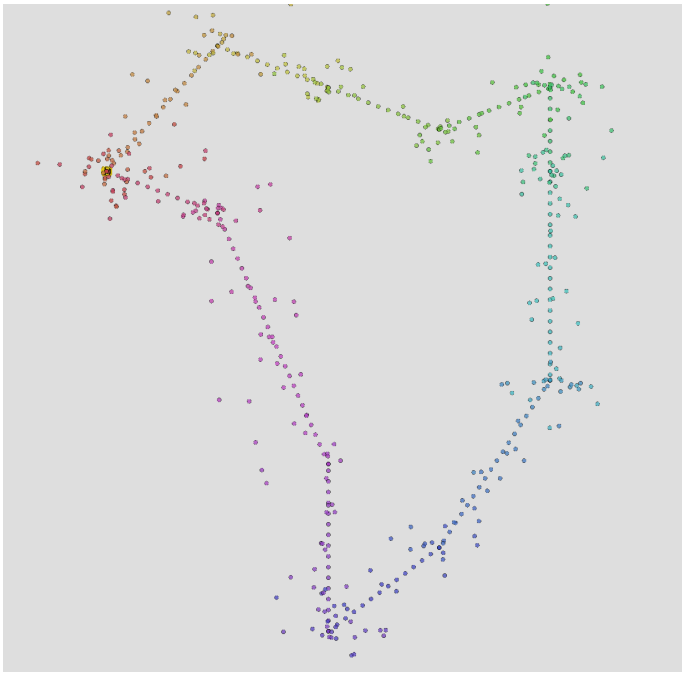

Figure 2: One agent original path overlaid with the Laplacian additive noise version

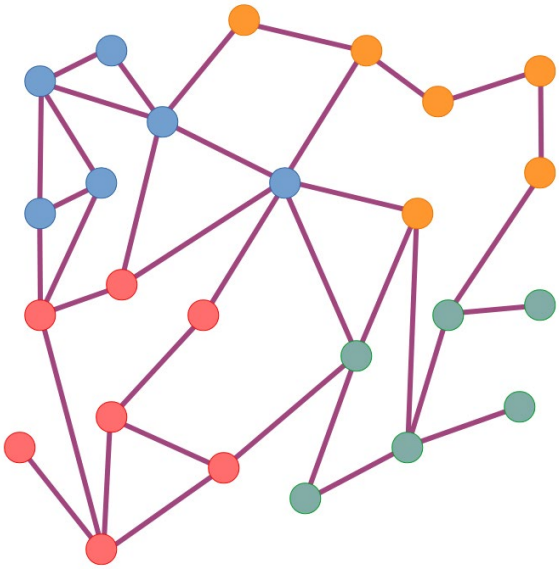

Figure 3: Example 2-dimensional graph used as a basis for spatiotemporal microdata streams

The synthesis tool also produced an alternate version of the CSV file with ε-DP perturbations applied independently to the timestamp and X and Y coordinate values using a Laplacian additive noise function and a configurable ε privacy loss budget, illustrated in Figure 3.

3.3 Adversarial Re-Identification

A separate artifact was developed to ingest the CSV format output by the synthesis tool to attempt to associate independent event data in differentially private outputs. This artifact was developed in .NET as a C# console application and used heuristic approaches to attempt re-association of the co-mingled agent microdata stream. Using synthesized data where original and ε-DP agent-tagged data were both available was vital in developing this artifact; however, it highlighted several methodological differences. In a small-world model graph, some associative strategies with significant adversarial value in the real world do not hold.

In particular, early research design cycles that assumed agents do not double back on their paths or are less likely to make sharp turns in succession do not hold when an A* path-finding algorithm is used among a small number of nodes. Further, as a model graph is a gross simplification of more intricate real-world

spatiotemporal microdata streams, heuristic approaches that do not encode hidden knowledge about the synthesis algorithm and fall off in accuracy quickly as the number of concurrent agents in the same location and timestamp increase due to a lack of defining attribute values over the period of cross-over in a small model space. Tuning the model to add nodes or create a variance in edge lengths and speeds can compensate in part for the observed challenges; however, real-world symmetrical spaces, such as dense city blocks with numerous one-way streets, similar path lengths, or multi-lane edges where agents could pass one another on a vector would exhibit similar limiting characteristics.

Because the model agents have only a constant speed and no acceleration, the associative strategy used for this adversarial artifact assumed agents are likely to travel in approximately the same direction, tolerating an error to account for ε-DP perturbations in sample placement. The heuristic was iteratively determined in the resulting checks:

- Discard events with more than (agentCount ^ 2) per 100 square unit lengths.
- For each remaining event, assume an associated event exists within 1.3 the maximum agent speed of the event sampling rate as measured as the distance between the (X,Y) coordinates for each.
- Assume three events must be within 1.3^2 the maximum agent speed

Check 1' aims to skip polynomial-time analysis of data points near nodes, where multiple agents may be waiting or intersecting to change directions. Success for the adversarial artifact is not gauged by absolute accuracy in associating all ε-DP events to their underlying agents but by tracing the path of individuals over the event microdata stream. In other words, a perfect association of timestamps and location at a virtual stoplight is unnecessary, provided the agents' behaviour can be determined once they leave an intersection.

Coefficient value 1.3 was arbitrarily chosen for the static model graph in Check 2, and every static exhibited drop-off effectiveness in association as it was highly dependent on the agent count, which affected the ε-DP sensitivity as well as the ε itself, which influenced the magnitude of perturbation if the Laplacian additive noise. The directional angle of candidate nodes was not a valuable heuristic consideration, as timestamp noise could reorder event stream sequences. Ordered sequences could still be perturbed out of spatial order at low ε values.

4. Results and Discussion

Three independent variables were considered to measure the accuracy of associating any given data point to its resulting agent set of data points: the privacy budget (ε), the number of agents, and the sampling rate of microdata. Higher values of ε result in lower privacy but higher utility as data points deviate less from their actual positions. As noted in Figures 4, 5, 6, and 7, the highest degrees of success for the adversarial artifact were, on average, for synthesis runs with the fewest agents and the highest sampling rates at every privacy-loss budget ε. As ε increases, so decreases the distance between ε-DP perturbed microdata event points, and more pronounced at higher agent counts. More agents intuitively can lower the effectiveness of the adversarial agent as there is a greater likelihood of multiple agents to be in the same approximately spatiotemporal location; however, more agents also decrease the sensitivity of the independent microdata attributes, resulting in a tighter bounding when Laplacian additive noise is applied. A lower sampling rate yields fewer data points and a higher degree of ambiguity, particularly on short paths where the angle of travel and approximate edge placement are indiscernible. Accuracy measured as an average for combinations of sampling rate and agent count is quantified in Table 1.

Table 1: Adversarial Artifact Associative Performance

Adversarial Artifact Associative Performance	
Privacy-loss budget	**Average Accuracy for Agent Count-Sampling Rate Combinations**
ε = 1	9%
ε = 3	25%
ε = 5	37%
ε = 10	52%

Notably, the adversarial artifact considers event density where agents may converge. It ignores those candidate events when attempting to reassociate points to agents, excluding likely node spatial locations from the accuracy measurement. For this reason, the high rates of accuracy at ε = 5 and ε = 10 in Figure 6 and Figure 7, respectively, are applicable only for events outside of the 100 square unit lengths density bounds as described in the heuristic methodology above. Additionally, because the differential privacy mechanism was naïvely applied to each tuple's attributes without frequent item-set protections previously identified in the literature by Zhen et al (2020) and others, this result alone does not represent a novel mechanism for reassociating spatiotemporal trajectories. However, it provides an artifact differential privacy implementers can use to analyse their outputs for errors in applying the approach.

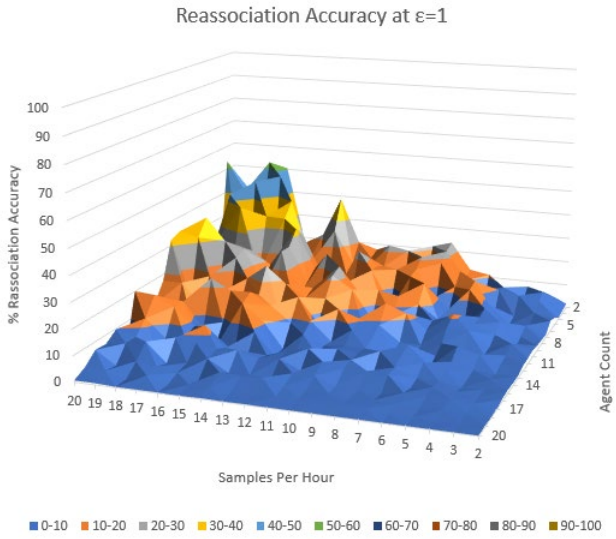

Figure 4: Weak associative performance at low ε=1 privacy-loss budget

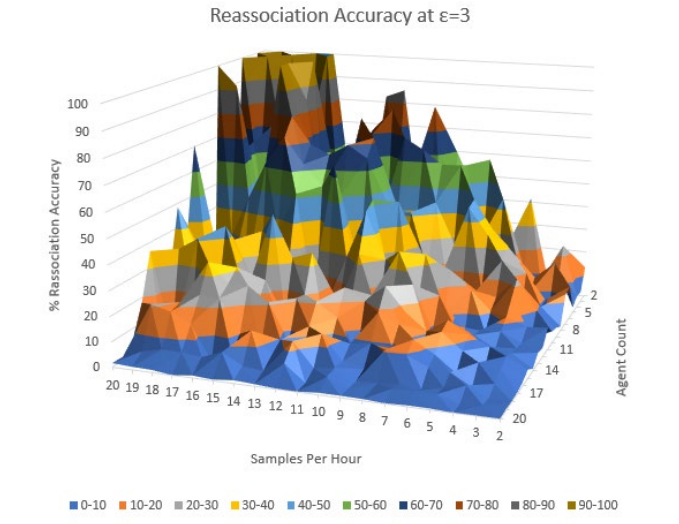

Figure 5: Improved associativity at extremities at higher ε=3

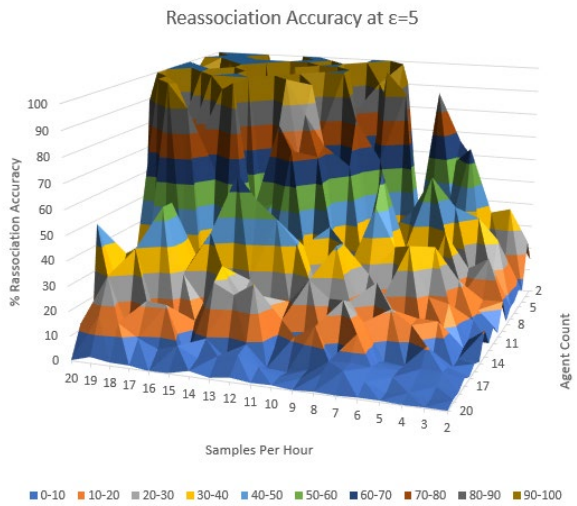

Figure 6: Frontier of high artifact performance markedly visible as ε increases to 5

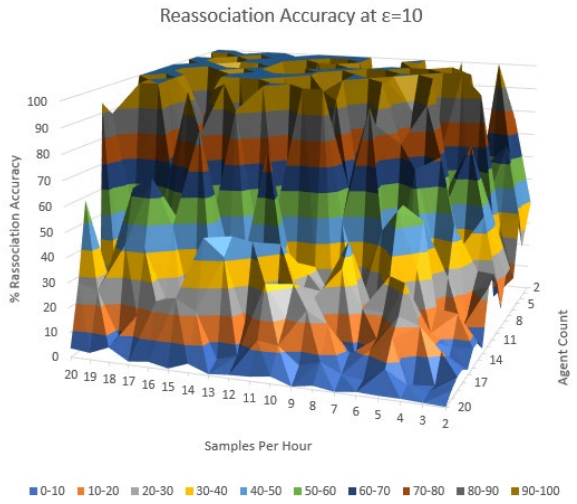

Figure 7: Frontier of high artifact performance as ε increases to 10 and privacy is traded-off for utility

5. Future Research

An artifact that can associate ε-DP spatiotemporal data serves as a guide for further research the approach to microdata streams that have robust, domain-specific approaches for protecting event-level microdata. Continuing technical action research to refine the adversarial artifact may yield additional heuristic or machine-learning approaches that improve performance and demonstrate associativity even with robust differential privacy approaches tailored for microdata streams. Future iterations of the artifact may yield more applications for validating differential privacy applications across event-level microdata and approaches for improving existing privacy-preserving techniques for highly correlated microdata, generally.

6. Conclusion

This technical action research yielded a design artifact that successfully associated spatiotemporal events microdata streams to individual agents in a small-world 2D planar graph simulating timestamped location coordinates with synthesized data. This adversarial artifact can be used to identify errors in the application of differential privacy through naïve means by attempting to circumvent privacy guarantees that misapplied differential privacy techniques fail to provide as desired through implementation error. The use of this artifact can help defend against advanced information warfare tactics targeting individuals based on their spatiotemporal patterns of life. Further research may enhance the artifact's ability to validate differentially private outputs for robust implementations of domain-specific, event-level techniques.

References

Agrawal, R., Imieliński, T., Swami, A. (1993) “Mining association rules between sets of items in large databases”, Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data, pp 207–216.

Dinur, I., Nissim, K. (2003) “Revealing Information While Preserving Privacy.” Proceedings of the Twenty-Second ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems, Association for Computing Machinery, pp 202–10.

Dwork, C. (2008) “Differential Privacy: A Survey of Results.” Theory and Applications of Models of Computation, Springer Verlag, Germany.

Dwork, C., Naor, M., Pitassi, T., Rothblum, G.N. (2010) "Differential Privacy under Continual Observation”, Proceedings of the Forty-Second ACM Symposium on Theory of Computing, pp 715–724.

Dwork, C., Roth, A. (2013) “The Algorithmic Foundations of Differential Privacy”, Frontiers in Theoretical Computer Science, Vol 9, No. 3-4, pp 211–407.

Hart, P., Nilsson, N., Raphael, B. (1968) “A Formal Basis for the Heuristic Determination of Minimum Cost Paths”, IEEE Transactions on Systems Science and Cybernetics, Vol 4, No. 2, pp 100–107.

Hoepman, J.-H. (2020) “Privacy Design Strategies (The Little Blue Book)”, [online], https://www.cs.ru.nl/~jhh/publications/pds-booklet.pdf.

Jarmin, R. (2019) “Census Bureau Adopts Cutting Edge Privacy Protections for 2020 Census”, [online], The United States Census Bureau, https://www.census.gov/newsroom/blogs/random-samplings/2019/02/census_bureau_adopts.html.

Li, N., Qardaji, W., Su, D., Cao, J. (2012) “PrivBasis: Frequent Itemset Mining with Differential Privacy”, Proceedings of the VLDB Endowment, Vol 5, No. 11, pp 1340–1351.

Wieringa, R.J. (2014) Design Science Methodology for Information Systems and Software Engineering, Springer Berlin Heidelberg, New York.

Xiao, Y., Xiong, L. (2015) “Protecting Locations with Differential Privacy under Temporal Correlations”, Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security, p. 1298–1309.

Xiao, Y., Xiong, L., Zhang, S., Cao, Y. (2017) “LocLok: Location Cloaking with Differential Privacy via Hidden Markov Model”, Proceedings of the VLDB Endowment, Vol 10, No. 12, pp 1901–1904.

Yilmaz, E., Ji, T., Ayday, E., Li, P. (2022) “Genomic Data Sharing under Dependent Local Differential Privacy”, Proceedings of the Twelfth ACM Conference on Data and Application Security and Privacy, pp 77–88.

Zhao, J., Mei, J., Matwin, S., Su, Y., Yang, Y. (2021) “Risk-Aware Individual Trajectory Data Publishing with Differential Privacy”, IEEE Access, Vol 9, pp 7421–7438.

Zhen, H., Chiou, B., Tsou, Y., Kuo, S., Wang, P. (2020) “Association Rule Mining with Differential Privacy”, 2020 50th Annual IEEE/IFIP International Conference on Dependable Systems and Networks Workshops, pp 47–54.