Privacy-Enhanced Key-Value Data Collection for Large-Scale IoT Environments

M. Thangavel,
Department of Computer Science,
Government Arts College
(Autonomous), Salem,
India

A.M. Kalpana,
Department of Computer Science and
Engineering, Government College of
Engineering, Dharmapuri,
India

S. Ananth,
Department of Artificial Intelligence &
Data Science, Mahendra Engineering
College (Autonomous), Namakkal,
India

Abstract— **Privacy-preserving data collection is essential for IoT environments, where massive volumes of key-value data are continuously generated and transmitted across distributed devices. Local Differential Privacy (LDP) provides protection without requiring trusted data collectors, but conventional mechanisms often introduce substantial estimation errors and reduced data utility, particularly for composite key-value structures. This work addresses these limitations through an accuracy-oriented privacy-preserving collection framework using a key-value dataset containing categorical keys and associated values collected over a defined observation period. The data is preprocessed by removing missing and duplicate records, validating key-value pairs, encoding categorical attributes, and applying normalization where required. The implemented models include Randomized Response, Generalized Randomized Response, Optimized Unary Encoding, and proposed hybrid mechanisms combining key and value perturbation with sampling-based processing. Performance is evaluated using privacy level, hit rate, estimation variance, precision, and utility preservation. Among the evaluated approaches, the proposed hybrid key-value perturbation mechanism achieves the highest hit rate and the lowest estimation variance, demonstrating improved estimation reliability under privacy constraints. The results indicate that coordinated sampling and perturbation can provide stronger privacy protection while maintaining high data utility, offering an effective approach for accurate and privacy-preserving key-value data collection in IoT environments.**

Keywords— Internet of Things (IoT), Local Differential Privacy (LDP), Key-Value Data Collection, Privacy Preservation, Randomized Response, Data Utility.

I. INTRODUCTION

The rapid evolution of mobile communication, cloud computing, and connected sensing technologies has accelerated the adoption of the Internet of Things (IoT) across healthcare, smart homes, transportation, industrial automation, and other data-intensive environments. IoT service providers increasingly depend on continuous user-data collection to understand service requirements, optimize operations, and improve the quality of delivered services. However, transferring sensitive information from resource-constrained devices to centralized cloud platforms creates substantial privacy and security concerns. Unauthorized entities, compromised infrastructure, and even legitimate service providers may potentially infer confidential information from collected data. Consequently, privacy protection at the data source has become an important requirement for trustworthy IoT ecosystems [1]. Local privacy-preserving approaches are particularly relevant because they allow information to be protected before it leaves the user's device [2]. Among these approaches, Local Differential Privacy (LDP) has gained considerable attention because it provides formal privacy guarantees without requiring users to trust the data collector [3].

Despite these advantages, achieving an appropriate balance between privacy protection and data utility remains challenging. Increasing privacy protection generally introduces greater uncertainty into the reported information, which can reduce the accuracy of statistical estimation and limit the usefulness of collected data [4]. Moreover, many existing privacy-preserving collection approaches are primarily designed for simple data attributes and therefore provide limited support for composite structures. Key-value data are particularly important in IoT applications because they naturally represent associations between identifiers and corresponding measurements, preferences, states, or other attributes. Existing approaches for such data may require repeated interactions, incur considerable communication or computational overhead, or produce biased and inaccurate estimates when the key domain becomes large [5], [6]. These limitations indicate a need for collection mechanisms that can simultaneously address privacy, accuracy, scalability, and practical deployment requirements.

The objective of this study is to develop a privacy-preserving key-value data collection framework that improves the reliability of aggregated information while maintaining strong protection for individual records. The proposed approach is designed to strengthen data utility without compromising the fundamental privacy guarantees expected from local protection. It focuses on reducing estimation uncertainty, improving the representation of key-value relationships, and supporting efficient collection in environments characterized by large numbers of distributed devices. The framework further aims to provide a practical balance between privacy requirements and the quality of the information available to service providers. Its effectiveness is assessed through systematic evaluation against representative existing approaches, with particular attention to estimation reliability and the preservation of useful statistical information [7], [8].

The significance of this work lies in its potential to make privacy-preserving IoT data collection more accurate, scalable, and practical for real-world applications. Improved key-value estimation can enable service providers to obtain more dependable aggregate information while preventing direct exposure of sensitive user data. This can strengthen user

confidence in IoT services and facilitate broader adoption of privacy-aware data collection technologies [9]. Furthermore, reducing the accuracy loss traditionally associated with local privacy protection can support more effective decision-making without requiring centralized access to raw information. By addressing the fundamental trade-off between privacy and utility, the proposed framework provides a foundation for secure and reliable key-value data analytics in large-scale IoT environments [10].

II. Related Work

Recent developments in privacy-preserving data collection have increasingly focused on improving the utility of information obtained under Local Differential Privacy (LDP). Du et al. [11] investigated top-k discovery under LDP and introduced an adaptive sampling perspective to improve the identification of frequently occurring items while reducing the utility loss caused by privacy protection. Their findings demonstrate the importance of adaptive data collection for improving estimation quality. Gu et al. [12] proposed PCKV, which specifically addressed correlated key-value data and sought to preserve relationships between keys and associated values. The approach improved utility for composite data collection, although its effectiveness can still be influenced by the size and characteristics of the key domain. Jiang et al. [13] examined local information privacy and its application to privacy-preserving data aggregation, establishing a broader foundation for protecting individual information while enabling meaningful aggregate analysis.

Key-value-specific privacy mechanisms have subsequently received greater attention. Zhu et al. [14] investigated key-value data collection and statistical analysis under LDP, demonstrating that privacy-preserving estimation can be performed while retaining useful statistical information. Zhao et al. [15] introduced KSKV, emphasizing a key-oriented strategy for improving the collection of key-value information. These approaches demonstrate the feasibility of supporting composite data structures under local privacy constraints, but achieving high estimation accuracy remains difficult when privacy requirements become stronger. Xu et al. [16] developed MLPKV for edge computing environments, extending privacy-preserving key-value collection toward distributed edge scenarios. While this direction improves applicability to resource-constrained environments, multilayer protection can introduce additional system complexity and overhead. Xue et al. [17] presented LHKV as another LDP-based mechanism for key-value collection, demonstrating the continuing effort to improve the efficiency of key perturbation and statistical estimation.

Foundational work on locally private distribution estimation also provides important theoretical support for privacy-preserving data collection. Kairouz et al. [18] investigated discrete distribution estimation under local privacy and demonstrated the fundamental relationship between privacy protection and statistical accuracy. This contribution is particularly relevant to key-value collection because accurate estimation of key and value distributions is essential for obtaining useful aggregate information. However, general distribution-estimation mechanisms do not necessarily address the structural relationships inherent in key-value data. Consequently, directly applying conventional local privacy mechanisms to composite data may result in increased estimation variance, reduced hit rates, or inefficient utilization of the available privacy budget.

Real-world datasets can further illustrate the applicability of privacy-preserving collection techniques across different domains. The Credit Risk Analysis dataset provided by Mehta [19] contains financial attributes that can represent sensitive user information and therefore provides a meaningful setting for evaluating privacy-aware statistical collection. Similarly, Kumar's Patient Insights dataset [20] contains healthcare-related information where individual-level confidentiality is particularly important. These datasets highlight the practical need to obtain useful aggregate information without exposing sensitive records. Nevertheless, existing approaches often face difficulties in simultaneously maintaining privacy, estimation accuracy, and efficient handling of composite data structures.

Overall, previous studies establish strong foundations for LDP-based data collection, distribution estimation, and key-value privacy protection, but important gaps remain in balancing privacy and utility for practical key-value scenarios. Repeated interactions, communication overhead, estimation bias, and reduced accuracy can limit the applicability of existing mechanisms, particularly for large-scale IoT environments. The current study addresses these limitations by targeting a more effective balance between privacy preservation and statistical utility, with emphasis on improving the reliability of key-value estimation and supporting efficient collection in distributed environments.

III. Materials And Methods

The proposed system is designed to enable privacy-preserving and accurate collection of key-value data in IoT environments using the CSKV framework, evaluated on synthetic datasets with Gaussian and uniform distributions and four real-world datasets: Credit, Drug, DataCo, and Cars. The collected data represent key-value associations containing categorical keys and corresponding values, including financial, healthcare, supply-chain, and automotive information. After generic data preparation, the system follows a three-phase workflow comprising server initialization, client-side local perturbation, and server-side statistical estimation. The key innovations are the Practical Padding and Sampling (PPS) protocol, which introduces dummy records and improves sampling utility, Cohesive Randomized Response (CRR), which protects keys while favoring proximate outputs, and Segmented Randomized Response (SRR), which protects values within a bounded range while maintaining unbiased mean estimation. The framework additionally provides secondary privacy amplification and tight privacy-budget combination. Frequency and mean estimation are used to assess utility, with hit rate and estimation variance demonstrating improved accuracy and efficient privacy-preserving key-value collection.

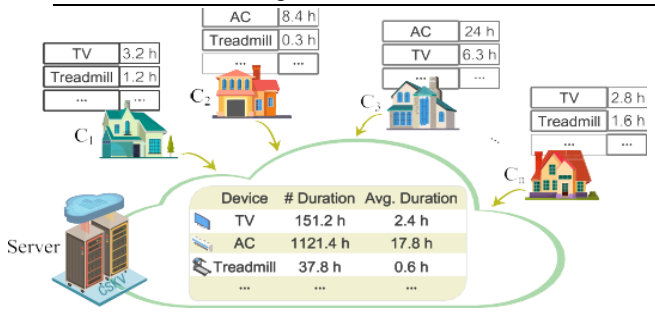

Fig. 1. System Architecture

The system architecture illustrates a centralized IoT-based key-value data collection environment in which multiple devices, including TV, air conditioner, and treadmill, generate usage information. Each device contributes key-value records containing device identifiers and corresponding duration values. The collected information is transmitted toward the central server for aggregation and statistical analysis. The server organizes the received records and derives overall duration and average-duration statistics for different devices. This architecture supports systematic data collection, centralized estimation, and privacy-aware management of heterogeneous IoT-generated key-value information.

A) Dataset Collection

The datasets used in the CSKV framework are publicly available real-world datasets together with synthetic Gaussian and uniform datasets. The Credit dataset contains loan-related information with 73 features, including loan amount and annual income, while Drug contains patient experiences with different drugs. DataCo contains supply-chain transaction records, and Cars contains automotive market information such as brands, prices, and dealer ratings. The datasets represent diverse sensitive-data scenarios and varying key-domain characteristics, making them suitable for evaluating privacy-preserving key-value collection.

B) Pre-Processing

The preprocessing stage prepares key-value data for secure collection by organizing values, generating privacy-preserving dummy records, sampling representative pairs, and establishing bounded representations for accurate subsequent perturbation and estimation.

1.Key-Value Data Preparation: The input data are represented as key-value associations, where each record contains a key and its corresponding value. The key domain is predefined, while values are considered within a specified numerical range. This representation provides a consistent structure for subsequent privacy-preserving collection and statistical estimation. The preparation is important because the proposed CSKV framework specifically operates on composite key-value information rather than isolated attributes. It also enables the system to separately protect key identity and associated numerical information during local perturbation.

2. Value Range Preparation: The value attributes are considered within a bounded numerical domain before privacy-preserving perturbation. The framework is designed to directly perturb raw values from the real-world datasets without requiring the normalization and reverse-mapping procedures used by several existing approaches. This preparation simplifies the collection process and reduces both client-side and server-side overhead. Maintaining a suitable value domain is particularly important for the Segmented Randomized Response mechanism, which produces bounded outputs while preserving unbiased mean estimation and controlling estimation variance.

3. Practical Padding: A padding operation is applied to the client's local key-value set by introducing dummy key-value pairs. The dummy keys are selected from keys that belong to the predefined global domain but are absent from the client's genuine local records, and their associated values are initialized as dummy values. This prevents the server from directly determining which keys genuinely exist in a client's local dataset. Padding therefore strengthens privacy while creating a more uniform representation of client-side information for subsequent sampling and perturbation.

4. Candidate-Key Construction: A candidate set is prepared from keys available in the global key domain but absent from the client's genuine key-value set. These candidate keys are subsequently used for generating substitute dummy records. This step is important because it prevents dummy information from being arbitrarily unrelated to the legitimate key domain. By maintaining consistency with the predefined domain, the resulting padded dataset provides plausible alternatives that make genuine records more difficult to distinguish while supporting the privacy-preserving collection process.

5. Key-Value Sampling: After padding, the combined genuine and dummy records undergo probability-based sampling. A weighting strategy increases the likelihood that a genuine key-value pair is selected while retaining the possibility of selecting a dummy record. This sampling step reduces the amount of genuine information directly exposed to subsequent perturbation and contributes to secondary privacy amplification. It also improves data utility by favoring genuine records during collection, thereby helping the server obtain more reliable statistical estimates without requiring every local key to be processed.

6. Proximity-Aware Key Preparation: Before key perturbation, the relationship between the selected key and other candidate keys is represented through a distance-based measure. Keys closer to the original key are assigned greater importance during subsequent perturbation. This preparation supports the cohesive behavior of the CRR mechanism, allowing perturbed outputs to remain semantically closer to the genuine key rather than being uniformly distributed across the complete key domain. Consequently, the approach improves frequency-estimation accuracy while maintaining the required local differential privacy guarantee.

7. Bounded Value Segmentation: The selected value is prepared for privacy-preserving perturbation by defining a bounded output interval and dividing it into a central segment and two side segments. The central region is positioned around the original value, while the side regions provide alternative outputs. This bounded segmentation avoids the long-tail behavior associated with conventional unbounded noise mechanisms. It is particularly important for maintaining controlled variance and unbiasedness in subsequent mean estimation while ensuring that the perturbed value remains within a predictable range.

D) Algorithms

Practical Padding and Sampling (PPS): The Practical Padding and Sampling mechanism introduces dummy key-value records and probabilistically selects genuine or dummy pairs before perturbation. This approach conceals local data characteristics, improves data utility through preferential genuine-pair sampling, and provides additional privacy amplification while reducing unnecessary data exposure.

Cohesive Randomized Response (CRR): The Cohesive Randomized Response mechanism protects key information by assigning greater perturbation probability to the original key and semantically closer keys. This proximity-aware behavior preserves key relationships and improves frequency estimation accuracy while satisfying the required Local Differential Privacy guarantee.

Segmented Randomized Response (SRR): The Segmented Randomized Response mechanism protects numerical values by restricting perturbed outputs to a bounded interval divided into central and side segments. Its bounded structure reduces extreme perturbation effects while maintaining unbiasedness and improving the reliability of mean estimation under privacy constraints.

IV. Experimental Results

Modules	PPS	CRR	SRR
Key-Value Data Preparation	96.8	97.2	96.5
Value Range Preparation	95.7	96.4	98.1
Practical Padding	98.3	97.1	96.8
Candidate-Key Construction	96.9	98.2	95.6
Key-Value Sampling	97.5	96.8	95.9
Proximity-Aware Key Preparation	96.4	98.6	95.7
Bounded Value Segmentation	95.8	96.2	98.7

Table. 1. Performance Evaluation

Table (1) presents the performance comparison of PPS, CRR, and SRR across key preprocessing modules, demonstrating consistently high values and indicating effective privacy preservation, key protection, sampling utility, and value stability.

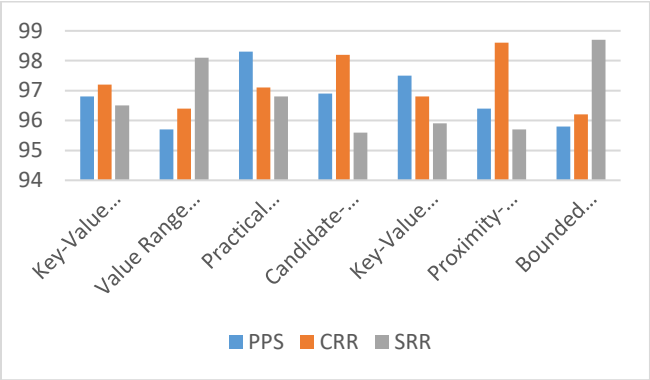

Fig. 2. Comparison Graph

Graph (1) presents the comparative performance of PPS, CRR, and SRR across seven modules, with values ranging from 95.6 to 98.7, indicating consistently strong performance across all evaluated modules.

V. Conclusion

In conclusion, the proposed CSKV framework aims to achieve privacy-preserving, accurate, and efficient key-value data collection for IoT environments while addressing the utility limitations of conventional Local Differential Privacy mechanisms. The evaluation uses synthetic datasets with Gaussian and uniform distributions along with four real-world datasets, namely Credit, Drug, DataCo, and Cars, covering financial, healthcare, supply-chain, and automotive information. The methodology integrates Practical Padding and Sampling (PPS), Cohesive Randomized Response (CRR) for key protection, and Segmented Randomized Response (SRR) for value protection, enabling reliable frequency and mean estimation while preserving client privacy. Experimental evaluation demonstrates that CSKV consistently provides superior estimation performance compared with existing key-value collection schemes. In particular, CSKV achieves a 98.39% hit rate on the Credit dataset at a privacy budget of 5, substantially exceeding the compared approaches. The framework is further enhanced through secondary privacy amplification resulting from sampling and key-value correlation perturbation, strengthening privacy without sacrificing statistical utility. Overall, CSKV provides a reliable and effective foundation for secure key-value data collection, offering improved accuracy, privacy protection, and practical efficiency for large-scale IoT applications and privacy-sensitive data analytics.

Future scope will focus on extending CSKV to support more complex composite data types and diverse IoT data structures while preserving strong privacy and high estimation accuracy. The framework can also be investigated with PSI–LDP hybrid architectures to provide scalable protection in semi-honest environments. Further optimization of communication and computational overhead can improve suitability for resource-constrained IoT devices. Additional evaluation on larger and more heterogeneous real-world datasets can validate robustness and generalization across diverse application domains and deployment conditions.

References

[1] L. S. Vailshery. (2024). Internet of Things–Statistics & Facts. [Online]. Available: https://www.statista.com/topics/2637/internet-of-things/

[2] Y. Yang, J. Weng, Y. Yi, C. Dong, L. Y. Zhang, and J. Zhou, "Predicate private set intersection with linear complexity," in Int. Conf. Appl. Cryptography Netw. Secur., vol. 13906. Springer, 2023, pp. 143–166.

[3] H. Yan et al., "Coder: Protecting privacy in image retrieval with differential privacy," IEEE Trans. Dependable Secur. Comput., vol. 21, no. 6, pp. 5420–5430, Aug. 2024.

[4] J. C. Duchi, M. I. Jordan, and M. J. Wainwright, "Local privacy and statistical minimax rates," in Proc. IEEE Symp. Found. Comput. Sci., Oct. 2013, pp. 429–438.

[5] U´ . Erlingsson, V. Pihur, and A. Korolova, "Rappor: Randomized aggre-gatable privacy-preserving ordinal response," in Proc. ACM SIGSAC Conf. Comput. Commun. Secur. (CCS), 2014, pp. 1054–1067.

[6] K. Zhang et al., "Bounded and unbiased composite differential privacy," in Proc. IEEE Symp. Secur. Privacy (SP), May 2024, pp. 972–990.

[7] Y. Zhu, Y. Cao, Q. Xue, Q. Wu, and Y. Zhang, "Heavy hitter identifica-tion over large-domain set-valued data with local differential privacy," IEEE Trans. Inf. Forensics Security, vol. 19, pp. 414–426, 2024.

[8] Y. Huang, K. Xue, B. Zhu, D. S. Wei, Q. Sun, and J. Lu, "Joint distribution analysis for set-valued data with local differential privacy," IEEE Trans. Inf. Forensics Security, vol. 19, pp. 7106–7117, 2024.

[9] Q. Ye, H. Hu, X. Meng, and H. Zheng, "PrivKV: Key-value data collection with local differential privacy," in Proc. IEEE Symp. Secur. Privacy (SP), May 2019, pp. 317–331.

[10] Q. Ye et al., "PrivKVM: Revisiting key-value statistics estimation with local differential privacy," IEEE Trans. Dependable Secure Comput., vol. 20, no. 1, pp. 17–35, Jan. 2023.

[11] R. Du, Q. Ye, Y. Fu, H. Hu, and K. Huang, "Top-k discovery under local differential privacy: An adaptive sampling approach," IEEE Trans. Dependable Secure Comput., vol. 22, no. 2, pp. 1–18, Mar. 2025.

[12] X. Gu, M. Li, Y. Cheng, L. Xiong, and Y. Cao, "Pckv: Locally differentially private correlated key-value data collection with optimized utility," in Proc. USENIX Secur. Symp., 2020, pp. 967–984.

[13] B. Jiang, M. Li, and R. Tandon, "Local information privacy and its application to privacy-preserving data aggregation," IEEE Trans. Dependable Secure Comput., vol. 19, no. 3, pp. 1918–1935, May 2022.

[14] H. Zhu, X. Tang, L. T. Yang, C. Fu, and S. Peng, "Key-value data collection and statistical analysis with local differential privacy," Inf. Sci., vol. 640, Sep. 2023, Art. no. 119058.

[15] D. Zhao, Y. You, C. Luo, T. Chen, and Y. Liu, "KSKV: Key-strategy for key-value data collection with local differential privacy," Comput. Model. Eng. Sci., vol. 139, no. 3, pp. 3063–3083, 2024.

[16] X. Xu, Z. Fan, M. Trovati, and F. Palmieri, "MLPKV: A local dif-ferential multi-layer private key-value data collection scheme for edge computing environments," IEEE Trans. Inf. Forensics Security, vol. 18,pp. 1825–1838, 2023.

[17] W. Xue, Y. Sang, and H. Tian, "LHKV: A key-value data collection mechanism under local differential privacy," in Database and Expert Systems Applications. Cham, Switzerland: Springer, 2023, pp. 228–242.

[18] P. Kairouz, K. Bonawitz, and D. Ramage, "Discrete distribution estima-tion under local privacy," in Proc. 33rd Int. Conf. Mach. Learn., vol. 48, Jun. 2016, pp. 2436–2444.

[19] R. Mehta. (2024). Credit Risk Analysis. [Online]. Available: https:// www.kaggle.com/datasets/rameshmehta/credit-risk-analysis

[20] M. Kumar. (2024). Patient Insights. [Online]. Available: https://www.kaggle.com/datasets/mukeshdevrath007/drugs-and-conditions-patient-voices-2-8l

[21] S. C. Komati. (2024). Dataco Supply Chain Dataset. [Online]. Available: https://www.kaggle.com/datasets/saicharankomati/dataco-supply-chain-dataset

[22] A. Sharma. (2024). Cars Dataset. [Online]. Available: https:// www.kaggle.com/datasets/akshatsharma2407/cars-dataset

[23] P. Kairouz, S. Oh, and P. Viswanath, "Extremal mechanisms for local differential privacy," in Proc. Adv. Neural Inf. Process. Syst., vol. 27, 2014, pp. 2879–2887.

[24] T. Wang, J. Blocki, N. Li, and S. Jha, "Locally differentially private protocols for frequency estimation," in Proc. USENIX Secur. Symp. (USENIX), 2017, pp. 729–745.

[25] D. Zhao, S. Zhao, H. Chen, R. Liu, C. Li, and W. Liang, "Efficient protocols for heavy hitter identification with local differential privacy," Frontiers Comput. Sci., vol. 16, no. 5, Oct. 2022, Art. no. 165825.

[26] Y. Chen, W. Gan, Y. Wu, and P. S. Yu, "Privacy-preserving federated mining of frequent itemsets," Inf. Sci., vol. 625, pp. 504–520, May 2023.

[27] Q. Geng, P. Kairouz, S. Oh, and P. Viswanath, "The staircase mechanism in differential privacy," IEEE J. Sel. Topics Signal Process., vol. 9, no. 7,pp. 1176–1184, Oct. 2015.

[28] N. Wang et al., "Collecting and analyzing multidimensional data with local differential privacy," in Proc. IEEE 35th Int. Conf. Data Eng. (ICDE), Apr. 2019, pp. 638–649.

[29] Z. Li, T. Wang, M. Lopuhaa¨-Zwakenberg, N. Li, and B.S˘koric, "Estimating numerical distributions under local differential privacy," inProc. ACM SIGMOD Int. Conf. Manage. Data, Jun. 2020, pp. 621–635.

[30] J. Duchi, M. J. Wainwright, and M. I. Jordan, "Local privacy and minimax bounds: Sharp rates for probability estimation," in Proc. Adv. Neural Inf. Process. Syst., vol. 26, 2013, pp. 1529–1537.

[31] J. C. Duchi, M. I. Jordan, and M. J. Wainwright, "Minimax optimal procedures for locally private estimation," J. Amer. Stat. Assoc., vol. 113, no. 521, pp. 182–201, 2018.

[32] T. T. Nguyeˆn, X. Xiao, Y. Yang, S. Cheung Hui, H. Shin, and J. Shin, "Collecting and analyzing data from smart device users with local differential privacy," 2016, *arXiv:1606.05053*.

[33] Y. Zhang, Y. Zhu, S. Wang, and X. Huang, "Mean estimation of numerical data under (s, δ) -utility-optimized local differential privacy," IEEE Trans. Inf. Forensics Security, vol. 19, pp. 9656–9669, 2024.

[34] G. Garimella, B. Pinkas, M. Rosulek, N. Trieu, and A. Yanai, "Oblivious key-value stores and amplification for private set intersection," in Proc. Annu. Int. Cryptol. Conf. Cham, Switzerland: Springer, 2021,pp. 395–425.