American Journal of Advanced Multidisciplinary Innovation and Research (AJAMIR)

E-ISSN: XXXX-XXXX ● Website: www.ajamir.com ● Email: editor@ajamir.com

Privacy-Preserving Federated Analytics for Collaborative Institutional Research

Dr. Matteo R. Bellini

Associate Professor, European Institute of Digital Governance, Milan, Italy

Abstract

Collaborative institutional research can generate stronger evidence about student progression, retention, curriculum effectiveness, access, academic support, and resource utilization than isolated analysis within individual institutions. Yet cross-institutional research frequently requires access to sensitive student-level records, creating substantial privacy, governance, security, and organizational barriers to conventional centralized data pooling. This paper examines privacy-preserving federated analytics as an alternative architecture for collaborative institutional research. Federated analytics enables participating institutions to perform approved computations locally and contribute restricted intermediate results to an aggregation process while retaining raw records within their own information environments.

The study adopts a conceptual-methodological research design grounded in contemporary developments in federated computation, secure aggregation, differential privacy, cross-silo learning, and privacy attacks. The proposed architecture combines local query execution, schema harmonization, contribution bounding, secure aggregation, minimum-participation thresholds, differential privacy, query authorization, privacy-budget management, audit logging, and disclosure review. Foundational secure-aggregation research demonstrates that distributed parties can compute aggregate values without revealing each participant's private input, while subsequent research shows that aggregated updates may still become vulnerable to reconstruction or disaggregation under inappropriate threat models.

Differential privacy provides an additional formal mechanism for limiting the influence of individual records on released statistics, although stronger privacy generally introduces a measurable utility cost. The analysis shows that federated institutional research is especially well suited to counts, proportions, means, histograms, frequency estimation, benchmark indicators, and selected statistical models, provided that query outputs are sufficiently aggregated and protected. The proposed framework also emphasizes that technical privacy mechanisms cannot substitute for common data definitions, institutional governance, purpose limitation, transparent participation agreements, bias assessment, and appropriate interpretation of heterogeneous institutional populations.

The paper concludes that federated analytics can support a transition from “share the data to conduct research” toward “share approved computations and protected aggregate evidence.” Its strongest value lies not in making sensitive data risk-free, but in reducing unnecessary centralization while preserving opportunities for collaborative, reproducible, and privacy-conscious institutional research.

American Journal of Advanced Multidisciplinary Innovation and Research (AJAMIR)

E-ISSN: XXXX-XXXX ● Website: www.ajamir.com ● Email: editor@ajamir.com

Keywords: Federated analytics; institutional research; privacy-preserving analytics; secure aggregation; differential privacy; higher education analytics; collaborative research; cross-silo analytics; data governance; student data privacy

1. Introduction

Institutional research increasingly depends on data that extend beyond a single university. Questions concerning student retention, transfer pathways, equity in academic progression, program effectiveness, learning-support utilization, graduate outcomes, and institutional benchmarking can require comparisons across organizations. Conventional collaborative research often addresses this requirement by transferring institution-level extracts or student-level datasets into a common repository. Although centralized pooling simplifies statistical analysis, it also expands the number of copies of sensitive data, increases the consequences of unauthorized access, requires complex governance arrangements, and can discourage institutions from participating when they cannot justify transferring raw records. The central methodological challenge is therefore to preserve the analytical benefits of multi-institution evidence without assuming that useful collaboration always requires centralized possession of the underlying records.

Federated analytics provides a different computational model. Google Research introduced the term in 2020 to describe the application of data-science methods to information retained locally, with local computations contributing only aggregate results rather than raw records. The principle is related to federate learning but differs in analytical purpose. Federated learning typically coordinates distributed model training, whereas federated analytics can address descriptive and statistical questions such as counts, sums, averages, histograms, quantiles, frequency estimation, and other aggregate queries. Contemporary production systems have expanded this concept considerably. The PAPAYA federated-analytics stack reported at NSDI 2025, for example, combines local transformations with privacy-preserving aggregation primitives and provides analyst-oriented query mechanisms while attempting to balance privacy, scalability, and practical usability.

This architecture is particularly relevant to higher-education institutional research because universities resemble a cross-silo environment. Each institution controls a comparatively large local dataset, possesses its own administrative and technical infrastructure, and may be willing to participate in collaborative research without transferring identifiable student records to peer institutions or a central research organization. In such a setting, an approved query could be distributed to participating universities; each institution would calculate a restricted local statistic; and an aggregation mechanism would return only an agreed collective result. For example, institutions could jointly estimate first-year progression rates across demographic or program categories without providing the coordinating analyst with individual student records. The principle resembles the broader distributed-computation architecture established by McMahan and colleagues, in which data remain decentralized while computation moves toward the data.

Keeping raw records local, however, should not be confused with eliminating privacy risk. Lam and colleagues demonstrated that repeatedly observed aggregated model updates can in some settings be disaggregated by an untrusted server, enabling attribution and subsequent inference concerning individual participants. Marchand and colleagues subsequently examined sample re-attribution attacks in

cross-silo settings protected by secure aggregation, further reinforcing the importance of explicitly defining the adversarial model rather than assuming that aggregation alone resolves confidentiality concerns. Similarly, secure aggregation protects individual contributions from direct inspection but does not automatically prevent sensitive information from being inferred from an excessively granular final aggregate. A complete privacy architecture must therefore control both what enters the aggregation protocol and what is ultimately released from it.

2. Objectives of the Study

2.1 To Develop a Federated Architecture for Institutional Research

The first objective is to develop an analytical architecture through which universities or comparable institutions can jointly calculate approved statistics without transferring their underlying student-level records to a common database. The proposed framework considers descriptive statistics, benchmark indicators, longitudinal summaries, frequency distributions, subgroup comparisons, and selected predictive or statistical analyses that can be expressed through local computations followed by protected aggregation.

2.2 To Integrate Multiple Layers of Privacy Protection

The second objective is to examine how local data retention, secure aggregation, differential privacy, contribution bounding, minimum cohort requirements, query restrictions, and disclosure review can operate as complementary controls. Secure aggregation is intended to prevent the coordinating service from seeing individual institutional contributions in clear form, while differential privacy is intended to limit the extent to which a released result reveals information attributable to any single protected record. Foundational research establishes both mechanisms as distinct tools with different privacy functions.

2.3 To Examine Utility, Governance, and Cross-Institution Heterogeneity

The third objective is to identify conditions under which privacy-preserving collaborative analysis remains scientifically useful. The study considers statistical utility, communication cost, privacy–accuracy trade-offs, non-identically distributed institutional populations, common variable definitions, auditability, fairness, and the need for explicit research governance.

3. Review of Literature

3.1 From Federated Learning to Federate Analytics

McMahan and colleagues established a practical federated-learning approach based on iterative model averaging across decentralized datasets. Their work showed that distributed learning could operate despite unbalanced and non-independent data distributions, establishing an important technical foundation for later federated computation. Federated analytics generalized the central idea beyond model training. Instead of requiring each party to contribute model updates, an analytics workflow can ask each participant to compute a limited local response to an approved query and then aggregate those responses.

The distinction is important for institutional research. Most university research offices do not need to train a deep-learning model to answer routine collaborative questions. They may instead require

statistics such as the number of students progressing to a second year, average credit accumulation, distributions of course completion, frequencies of academic-support use, or cross-institution benchmark indicators. Federated analytics can potentially answer such questions with fewer communication rounds and a narrower disclosure surface than conventional model-training workflows.

Production-scale developments demonstrate that federated analytics can support more expressive workflows. The PAPAYA system described by Srinivas and colleagues supports local data transformations, secure-sum primitives, thresholding, and analyst-oriented query specification. Its architecture separates device-side computation, trusted processing components, and orchestration, illustrating that practical federated analytics requires coordinated engineering rather than a single privacy algorithm. Although PAPAYA is primarily designed for cross-device use, the architectural principle is transferable to cross-institution research: local execution should be separated from protected aggregation and from the analyst-facing query layer.

3.2 Secure Aggregation and Differential Privacy

Secure aggregation addresses a fundamental problem in collaborative analytics: a central coordinator may need the total or average of local contributions without being permitted to inspect each contribution separately. Bonawitz and colleagues designed a communication-efficient protocol through which mutually distrustful participants can collaborate to compute an aggregate while withholding individual private inputs from the aggregator. Their protocol was designed to tolerate substantial participant dropout and demonstrated practical communication costs for high-dimensional inputs.

However, cryptographic aggregation protects the inputs to a calculation rather than automatically protecting the information contained in the final result. Consider a federation in which only two institutions contribute to a highly specific subgroup query. Even if their local counts are cryptographically hidden, the released total combined with outside knowledge could reveal information about one institution's contribution. Differential privacy addresses a different layer of this problem by limiting how strongly the presence or absence of an individual record can influence a released output. Dwork's foundational formulation provided a mathematical basis for privacy-preserving statistical analysis, while subsequent federated research has combined differential privacy with secure aggregation to obtain both protected computation and formal output privacy.

3.3 Privacy Threats, Heterogeneity, and Cross-Silo Collaboration

Federation reduces data movement but does not remove adversarial behavior. Lam and colleagues showed that summary information concerning participation could assist an untrusted server in reconstructing the participant matrix and disaggregating contributions from repeated aggregate observations. Their experiments demonstrated attacks at scales involving thousands of participants. This finding has direct implications for institutional research: query histories, participation metadata, subgroup structures, and repeated releases may collectively expose more information than any single query.

Cross-silo settings also differ from large cross-device federations because they contain fewer participants, while each institution may represent many individuals. Research by Lowy and colleagues examines private cross-silo federated learning when people represented within institutional silos do not

trust the coordinating server or other silos, demonstrating why record-level protection can be more relevant than merely hiding which institution participated. Similarly, Bietti and colleagues demonstrate that personalization and privacy interact in federated settings, while Cyffers and Bellet show that decentralization itself can modify privacy–utility trade-offs.

4. Research Methodology

4.1 Research Design

The study uses a conceptual-methodological and architecture-development design. No university-level student dataset was collected or simulated as though it represented actual institutional outcomes. Instead, the analysis synthesizes validated principles from secure multiparty computation, federated analytics, differential privacy, cross-silo federated computation, and documented privacy attacks to construct research architecture suitable for higher-education collaboration.

The proposed model assumes a federation of autonomous institutions. Each institution retains its source records locally, maps approved variables to a shared research schema, executes locally authorized computations, and returns only cryptographically protected intermediate values. A coordinating service is responsible for orchestration but is not assumed to be entitled to inspect raw institutional contributions.

4.2 Proposed Federated Institutional Research Workflow

The workflow begins with a jointly approved research question and standardized variable specification. Local institutional nodes then calculate bounded contributions. These contributions enter secure aggregation, after which minimum-participation rules and differential-privacy mechanisms are applied before an analyst receives the final statistic.

Table 1: Proposed Privacy-Preserving Federated Institutional Research Architecture

Stage	Institutional Operation	Privacy / Security Control	Research Output
Research authorization	Define purpose, variables, population, and analysis	Query approval and purpose limitation	Approved federated query
Schema harmonization	Map local fields to common definitions	No raw-data transfer	Comparable local variables
Local computation	Calculate counts, sums, moments, or model statistics	Data remain within institution	Local intermediate result
Contribution control	Bound values and suppress invalid categories	Contribution limits	Restricted local vector
Protected transmission	Encrypt or mask local contribution	Secure aggregation	Unreadable individual contribution
Aggregate computation	Combine authorized contributions	Minimum cohort/participant	Collective aggregate

American Journal of Advanced Multidisciplinary Innovation and Research (AJAMIR)

E-ISSN: XXXX-XXXX ● Website: www.ajamir.com ● Email: editor@ajamir.com

		threshold	
Privacy protection	Add calibrated statistical noise where required	Differential privacy	Privacy-protected statistic
Disclosure review	Assess small cells, differencing, repeated queries	Privacy-budget and query-history controls	Approved release
Analyst access	Provide aggregate research result	Role-based access and audit logging	Collaborative evidence
Reproducibility	Store query definition and metadata	Controlled provenance record	Auditable research workflow

4.3 Evaluation Criteria

A real deployment of this framework should be evaluated across five dimensions:

Privacy, statistical utility, communication efficiency, reproducibility, and governance reliability.

Privacy evaluation should specify the adversarial model and quantify differential-privacy parameters where DP is used. Utility should measure the difference between protected and unprotected aggregate estimates rather than assuming that privacy noise is negligible. Communication evaluation should include protocol overhead, response latency, participant dropout, and repeated-query cost. Reproducibility requires common query definitions, versioned schemas, and recorded privacy parameters. Governance evaluation should determine whether the released analysis corresponds strictly with the approved research purpose.

5. Analysis

5.1 Suitable Institutional Research Queries

Federated analytics is not equally appropriate for every institutional research problem. Its clearest use cases are analyses that can be decomposed into local sufficient statistics and recombined into a valid global estimate.

Table 2: Institutional Research Applications and Recommended Privacy Controls

Research Question	Federated Statistic	Principal Privacy Risk	Recommended Control
Multi-institution enrollment profile	Counts / proportions	Small-cell identification	Secure sum + threshold + DP
First-year persistence benchmarking	Counts / rates	Institutional contribution inference	Secure aggregation + minimum federation size
Mean credit accumulation	Sum + count	Extreme-value influence	Contribution clipping + secure aggregation
Course-outcome distribution	Histogram	Sparse category disclosure	Thresholded DP histogram

American Journal of Advanced Multidisciplinary Innovation and Research (AJAMIR)

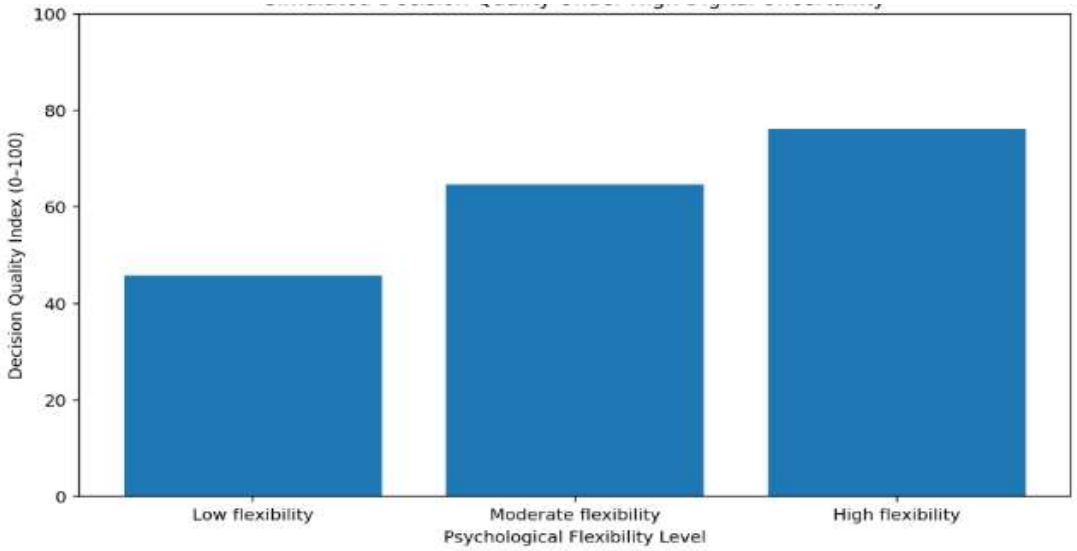
E-ISSN: XXXX-XXXX ● Website: www.ajamir.com ● Email: editor@ajamir.com

Support-service utilization	Frequency estimate	Sensitive subgroup exposure	DP frequency estimation
Equity-gap comparison	Stratified rates	Intersectional small cells	Restricted dimensions + threshold + DP
Cross-institution regression	Local sufficient statistics	Repeated-query leakage	Protected aggregation + privacy accounting
Predictive research	Federated model updates	Gradient/model inversion	Secure aggregation + DP + attack testing
Common emerging category discovery	Private heavy hitters	Rare-item disclosure	DP heavy-hitter protocol
Longitudinal benchmark monitoring	Repeated aggregates	Differencing and composition attacks	Cumulative privacy budget + release scheduling

5.2 Communication Cost of Privacy Protection

Cryptographic protection has a computational and communication cost. Bonawitz and colleagues reported that their practical secure-aggregation protocol, using 16-bit values, produced approximately 1.73× communication expansion for 1,024 participants and 1,048,576-dimensional vectors and 1.98× expansion for 16,384 participants and 16,777,216-dimensional vectors. The protocol was designed to tolerate failure of up to one-third of participating users.

Figure 1: Communication Expansion in a Practical Secure-Aggregation Protocol



Interpretation: For institutional research, the precise overhead will differ because the number of participating universities is generally much smaller and descriptive query vectors may be far lower dimensional. The benchmark should therefore be interpreted as evidence of protocol feasibility rather than as a direct estimate of university deployment cost.

American Journal of Advanced Multidisciplinary Innovation and Research (AJAMIR)

E-ISSN: XXXX-XXXX ● Website: www.ajamir.com ● Email: editor@ajamir.com

6. Discussion

6.1 From Data Sharing to Computation Sharing

The principal conceptual contribution of federated institutional research is a change in the object being exchanged. Traditional collaboration asks institutions to share datasets. The proposed model asks them to execute standardized computations and contribute protected results.

This distinction can reduce the proliferation of duplicate sensitive datasets and allows participating universities to retain operational control over their local records. It also permits institutions to verify query definitions before execution. Federated analytics was explicitly introduced as a means of conducting data science without collecting raw information centrally, and modern systems demonstrate that the model can support increasingly expressive analytical tasks.

For institutional researchers, the architecture could support consortium benchmarking, multi-campus equity analysis, coordinated evaluation of student-support interventions, and comparative progression studies while reducing the need to construct a permanent central student-data warehouse.

6.2 Why Federation Alone Is Insufficient

A recurring misconception is that data are automatically private whenever they remain local. The attack literature makes this assumption difficult to defend. Lam and colleagues demonstrated that aggregated updates may be disaggregated when repeated observations and participation information are available. Other attack research has shown that malicious servers can manipulate federated workflows to recover sensitive information from client updates. The practical consequence is that system designers must ask not only, “Does raw data leave the institution?” but also, “What can an adversary infer from the sequence of intermediate and final outputs?”

Institutional research introduces additional risks because institutions may be few in number. A federation of five universities does not provide the same crowd size as an aggregation involving thousands of devices. For this reason, minimum-participation thresholds, record-level differential privacy, query-history controls, and restrictions on highly granular cross-tabulations are particularly important.

Differential privacy itself introduces a utility cost. Kairouz and colleagues demonstrate that private distributed aggregation involves a trade-off among privacy, accuracy, communication, and quantization. Privacy parameters should therefore be treated as scientific design choices and documented in research outputs rather than hidden as purely technical implementation details.

6.3 Institutional Heterogeneity and Research Validity

Privacy-preserving computation solves only one class of collaborative-research problems. It does not guarantee that the resulting statistic has a coherent meaning.

Consider a federation comparing “first-year retention.” One institution may define retention as enrollment on a census date one year later, another may count any enrollment during the following academic year, and a third may classify internal transfer differently. Securely aggregating these values would produce a mathematically precise but conceptually invalid statistic.

The proposed framework therefore places schema harmonization before cryptographic aggregation. Participating institutions should agree on populations, exclusion rules, time windows,

missing-data treatment, coding systems, derived variables, and permitted subgroup dimensions before executing a federated query.

Data heterogeneity also affects statistical modeling. McMahan and colleagues identified non-independent and unbalanced data distributions as characteristic challenges of federated computation, and later cross-silo research has continued to address divergence among institutional data distributions. In higher education, such heterogeneity may be substantively meaningful rather than merely statistical noise. Institutional differences should therefore sometimes be modeled explicitly rather than averaged away.

Finally, federated research should preserve interpretability. A consortium should be able to determine which institutions participated, which common definitions were used, what privacy protection was applied, what uncertainty was introduced by DP noise, and which queries preceded the published result. Privacy preservation should strengthen responsible research practice rather than make analytical provenance opaque.

7. Conclusion

Privacy-preserving federated analytics offers a technically credible and methodologically promising alternative to conventional centralized data pooling for collaborative institutional research. By moving approved computations toward distributed institutional datasets and returning only protected aggregate information, universities can investigate shared research questions while reducing unnecessary transfer and duplication of student-level records.

The present study shows, however, that federated does not automatically mean private. Secure aggregation can conceal institutional contributions during computation, but the final analytical result may remain disclosure-prone. Differential privacy can provide formal protection at the output layer, but introduces privacy–utility trade-offs. Repeated queries can accumulate disclosure risk, while data heterogeneity and incompatible institutional definitions can undermine validity even when the computation is cryptographically secure.

The proposed architecture therefore combines local data retention, harmonized schemas, bounded contributions, secure aggregation, minimum participation thresholds, differential privacy, privacy-budget management, query authorization, audit logs, and disclosure review. It is particularly suitable for aggregate institutional statistics, benchmarking, subgroup comparisons, frequency analysis, and selected statistical modeling.

Future empirical work should establish test federations among consenting institutions and evaluate privacy loss, aggregate accuracy, communication overhead, analyst usability, schema interoperability, and governance performance under realistic institutional conditions. Research should also investigate how differential-privacy parameters affect small institutional subgroups and whether privacy protections introduce unequal statistical error across populations.

The long-term contribution of federated analytics is therefore not the elimination of privacy risk. It’s more defensible contribution is data minimization by architecture: enabling institutions to collaborate on evidence while revealing only the information required answering an authorized research question.

References

1. B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, "Communication-efficient learning of deep networks from decentralized data," in *Proc. 20th Int. Conf. Artificial Intelligence and Statistics (AISTATS)*, PMLR, vol. 54, pp. 1273–1282, 2017.
2. C. Dwork, "Differential privacy," in *Proc. 33rd International Colloquium on Automata, Languages and Programming*, 2006, pp. 1–12.
3. K. Bonawitz, V. Ivanov, B. Kreuter, A. Marcedone, H. B. McMahan, S. Patel, D. Ramage, A. Segal, and K. Seth, "Practical secure aggregation for privacy-preserving machine learning," in *Proc. ACM SIGSAC Conf. Computer and Communications Security*, 2017.
4. P. Kairouz, Z. Liu, and T. Steinke, "The distributed discrete Gaussian mechanism for federated learning with secure aggregation," in *Proc. 38th Int. Conf. Machine Learning*, PMLR, vol. 139, pp. 5201–5212, 2021.
5. M. Lam, G.-Y. Wei, D. Brooks, V. J. Reddi, and M. Mitzenmacher, "Gradient disaggregation: Breaking privacy in federated learning by reconstructing the user participant matrix," in *Proc. 38th Int. Conf. Machine Learning*, PMLR, vol. 139, pp. 5959–5968, 2021.
6. W.-N. Chen, C. A. Choquette-Choo, P. Kairouz, and A. T. Suresh, "The fundamental price of secure aggregation in differentially private federated learning," in *Proc. 39th Int. Conf. Machine Learning*, PMLR, vol. 162, pp. 3056–3089, 2022.
7. X. Zhang et al., "Understanding clipping for federated learning: Convergence and client-level differential privacy," in *Proc. 39th Int. Conf. Machine Learning*, PMLR, vol. 162, 2022.
8. Bietti, C. Wei, M. Dudík, J. Langford, and S. Wu, "Personalization improves privacy–accuracy tradeoffs in federated learning," in *Proc. 39th Int. Conf. Machine Learning*, PMLR, vol. 162, 2022.
9. E. Cyffers and A. Bellet, "Privacy amplification by decentralization," in *Proc. 25th Int. Conf. Artificial Intelligence and Statistics*, PMLR, vol. 151, pp. 5334–5353, 2022.
10. Lowy, A. Ghafelebashi, and M. Razaviyayn, "Private non-convex federated learning without a trusted server," in *Proc. 26th Int. Conf. Artificial Intelligence and Statistics*, PMLR, 2023.
11. T. Marchand, R. Loeb, U. Marteau-Ferey, J. Ogier du Terrail, and A. Pignet, "SRATTA: Sample re-attribution attack of secure aggregation in federated learning," in *Proc. 40th Int. Conf. Machine Learning*, PMLR, vol. 202, 2023.
12. W. Zhu, P. Kairouz, B. McMahan, H. Sun, and W. Li, "Federated heavy hitters discovery with differential privacy," in *Proc. 23rd Int. Conf. Artificial Intelligence and Statistics*, PMLR, pp. 3837–3847, 2020.
13. H. Srinivas, G. Cormode, M. Honarkhah, S. Lurye, J. Hehir, L. He, G. Hong, A. Magdy, D. Huba, K. Wang, S. Guo, and S. Bhattacharya, "PAPAYA federated analytics stack: Engineering privacy, scalability and practicality," in *Proc. 22nd USENIX Symposium on Networked Systems Design and Implementation (NSDI)*, pp. 883–898, 2025.
14. Z. Li et al., "APPFLx: Providing privacy-preserving cross-silo federated learning as a service," in *Proc. IEEE International Conference on eScience*, 2023.
15. J. Nguyen et al., "Federated learning with buffered asynchronous aggregation," in *Proc. 25th Int. Conf. Artificial Intelligence and Statistics*, PMLR, vol. 151, 2022.

16. W. Bao et al., "Optimizing the collaboration structure in cross-silo federated learning," in *Proc. 40th Int. Conf. Machine Learning*, PMLR, vol. 202, 2023.
17. E. Bagdasaryan et al., "How to backdoor federated learning," in *Proc. 23rd Int. Conf. Artificial Intelligence and Statistics*, PMLR, vol. 108, 2020.
18. Y. J. Cho et al., "Towards understanding biased client selection in federated learning," in *Proc. 25th Int. Conf. Artificial Intelligence and Statistics*, PMLR, vol. 151, 2022.