IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)
An International Open Access, Peer-reviewed, Refereed Journal

Privacy-Preserving Data Analysis: A Survey

Mr.Manjunath.N.M
MTech(Student), MVJ College of Engineering, Bangalore

Prof.Hetal Rana
Assistant Professor, MVJ College of Engineering, Bangalore, India

Abstract:

Privacy-preserving data analysis has emerged as a crucial area of research in response to the increasing concerns about the privacy of sensitive information in data-driven applications. This paper presents a comprehensive survey of techniques and methodologies in privacy-preserving data analysis. It covers various approaches, including differential privacy, homomorphic encryption, secure multiparty computation, and federated learning, highlighting their strengths, limitations, and applications. Additionally, we discuss open challenges and future directions in the field.

Keywords: Privacy, Data Analysis, Differential Privacy, Homomorphic Encryption, Secure Multiparty Computation, Federated Learning.

I. Introduction:

In the era of big data, the proliferation of data collection and analysis has raised significant concerns about individual privacy. Organizations collect vast amounts of data from individuals, including personal, financial, and health-related information, to derive valuable insights and improve decision-making processes. However, the unauthorized disclosure or misuse of sensitive data can lead to severe consequences, such as identity theft, financial fraud, and discrimination.

Privacy-preserving data analysis aims to enable organizations to extract valuable insights from data while protecting the privacy of individuals. It encompasses a range of techniques and methodologies that allow data analysis to be performed without revealing sensitive information. In this survey, we provide an overview of various privacy-preserving techniques and their applications in data analysis.

II. Differential Privacy:

Differential privacy, a cornerstone of modern privacy-preserving data analysis, provides a robust framework for ensuring privacy in data analysis tasks. It achieves this by introducing controlled noise or randomness into the computation process, thereby preventing adversaries from deducing sensitive information about individuals from the outputs of analyses. Originally proposed by Cynthia Dwork and colleagues in 2006, differential privacy has gained significant attention due to its strong mathematical guarantees and versatility in various applications.

Differential privacy is a rigorous privacy framework that provides strong privacy guarantees by ensuring that the presence or absence of an individual's data does not significantly affect the outcome of the analysis. It introduces randomness into the data analysis process to prevent adversaries from inferring sensitive information about individuals. Differential privacy has been applied to various data analysis tasks, including query answering, machine learning, and data publishing.

III. Homomorphic Encryption:

Homomorphic encryption stands as a formidable pillar in the realm of privacy-preserving data analysis. This cryptographic technique enables computations to be performed on encrypted data without the need for decryption, thereby preserving the confidentiality of sensitive information throughout the computation process. Developed as early as the late 1970s, homomorphic encryption has evolved into a sophisticated tool, offering a balance between privacy and utility in scenarios where data must be outsourced for analysis while maintaining strict privacy constraints.

Homomorphic encryption is a cryptographic technique that allows computations to be performed on encrypted data without decrypting it first. It enables privacy-preserving data analysis by allowing data to be securely outsourced to third-party servers for computation while ensuring that the data remains encrypted throughout the process. Homomorphic encryption has applications in secure cloud computing and outsourced data analysis.

IV. Secure Multiparty Computation:

Secure multiparty computation (SMC) offers a powerful mechanism for enabling collaborative data analysis while preserving the privacy of individual inputs. Originating from the seminal work of Andrew Yao in the early 1980s, SMC protocols allow multiple parties to jointly compute a function over their private inputs without revealing anything beyond the output of the computation. This decentralized approach to privacy-preserving data analysis finds applications in domains where data is distributed across multiple entities, such as health-care consortiums and collaborative research initiatives.

Secure multiparty computation (SMC) allows multiple parties to jointly compute a function over their inputs while keeping their inputs private. It enables privacy-preserving data analysis in scenarios where data is distributed across multiple parties who do not fully trust each other. SMC protocols ensure that each party learns only the output of the computation and nothing about the inputs of the other parties.

V. Federated Learning:

Federated learning, a novel paradigm in machine learning, revolutionizes the landscape of privacy-preserving data analysis. Introduced by Google researchers in 2016, federated learning enables model training across decentralized devices or servers while keeping raw data localized and private. By aggregating model updates instead of raw data, federated learning mitigates privacy risks associated with centralized data aggregation, making it particularly suitable for applications in sensitive domains like health-care and finance.

Federated learning is a decentralized approach to machine learning where the model is trained across multiple devices or servers holding local data samples, without exchanging them. It allows privacy-preserving data analysis by ensuring that the raw data never leaves the local device, and only model updates are shared with a central server. Federated learning has applications in various domains, including health-care, finance, and telecommunications.

VI. Challenges and Future Directions:

While privacy-preserving data analysis has made significant strides in recent years, several challenges remain. These include scalability issues, usability concerns, and the need for robust privacy guarantees in real-world applications. Addressing these challenges requires interdisciplinary research efforts spanning computer science, mathematics, and social sciences. Future directions in the field include developing more efficient and scalable privacy-preserving techniques, integrating privacy into the entire data analysis pipeline, and addressing emerging privacy threats such as membership inference and model inversion attacks.

VII. Conclusion:

Privacy-preserving data analysis plays a crucial role in ensuring the privacy and security of sensitive information in data-driven applications. In this survey, we have provided an overview of various privacy-preserving techniques and their applications in data analysis. While significant progress has been made in the field, there remain several challenges and open research directions that warrant further investigation. By addressing these challenges, we can pave the way for the development of more privacy-preserving and trustworthy data analysis techniques in the future.

References:

[1] Dwork, C., & Roth, A. (2014). The algorithmic foundations of differential privacy. Foundations and Trends® in Theoretical Computer Science, 9(3-4), 211-407.

[2] Gentry, C. (2009). A fully homomorphic encryption scheme. PhD thesis, Stanford University.

[3] Yao, A. C. (1982). Protocols for secure computations. In Proceedings of the 23rd Annual IEEE Symposium on Foundations of Computer Science (pp. 160-164). IEEE.

[4] Konečný, J., McMahan, H. B., Yu, F. X., Richtárik, P., Suresh, A. T., & Bacon, D. (2016). Federated learning: Strategies for improving communication efficiency. arXiv preprint arXiv:1610.05492.

[5] Abadi, M., Chu, A., Goodfellow, I., McMahan, H. B., Mironov, I., Talwar, K., & Zhang, L. (2016). Deep learning with differential privacy. In Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security (pp. 308-318). ACM.

[6] Boneh, D., & Shacham, H. (2004). Group signatures with verifier-local revocation. In Proceedings of the 11th ACM Conference on Computer and Communications Security (pp. 168-177). ACM.

[7] Kairouz, P., McMahan, H. B., Avent, B., Bellet, A., Bennis, M., Bhagoji, A. N., ... & Song, D. (2019). Advances and open problems in federated learning. arXiv preprint arXiv:1912.04977.

[8] Lauter, K., López-Alt, A., & Naehrig, M. (2011). Private computation on encrypted genomic data. In Proceedings of the 2011 ACM Workshop on Cloud computing security (pp. 33-44). ACM.

[9] Mohassel, P., & Zhang, Y. (2017). SecureML: A system for scalable privacy-preserving machine learning. In Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security (pp. 703-720). ACM.

[10] Shokri, R., & Shmatikov, V. (2015). Privacy-preserving deep learning. In Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security (pp. 1310-1321). ACM.