WeStat: A Privacy-Preserving Mobile Data Usage Statistics System

Shubhika GARG
Bipin ADHIKARI
Mayank NARANG
Lohith RACHAKONDA

I. SUMMARY OF THE WORK

The research paper explores the challenges and opportunities arising from the widespread use of mobile applications, resulting in a significant increase in mobile data. Recognizing the integral role of smartphones in daily life, the study highlights the potential to understand user behavior through app usage data, emphasizing the sensitivity of such data and the ensuing privacy concerns during collection and analysis.

The practical use case presented focuses on social scientists studying user habits, particularly within social network applications. Criticizing traditional survey methods like face-to-face interviews, diaries, questionnaires, or surveys for their inherent biases, the paper introduces WeStat as an alternative. WeStat not only mitigates these concerns but also ensures participant privacy during data analytics. Through the use of functional encryption, the approach enables statistical operations on encrypted data, facilitating robust analytics without compromising individual privacy.

A. *Proposed Solution*

The solution involves encrypting users' mobile usage data with a study-specific key on their devices. The Aggregator, acting as the Service Provider, performs analytical operations on the encrypted data and supplies results to the Third Party for decryption. Centered on simple analytics such as computing the mean, this approach adapts the DSum technique—a form of dynamic decentralized multi-client Functional Encryption.

Moreover, the fault-tolerant Private Stream Aggregation concepts are seamlessly integrated to accommodate dynamic user participation. This requires thoughtful adjustments to the DSum construction, ensuring the scalability and adaptability of the solution to varying user engagement levels. The All-or-Nothing encapsulation technique is strategically employed twice, while randomness on the Third-Party side prevents the Aggregator from gaining insights into partial results, ensuring a robust and privacy-preserving analytics process.

B. *Emphasis on Privacy Regulations and GDPR Compliance*

The paper consistently underscores strict privacy rules and the imperative need to adhere to the General Data Protection Regulation (GDPR) [1]. The system is meticulously designed to secure clear user agreement before data collection, utilizing advanced technical solutions, particularly encryption methods, to minimize identification risks. Notably, the Aggregator, handling encrypted data, is restricted from accessing personal information, and the Third Party, crucial for analytics, only receives final results, maintaining a non-interactive role in data processing. This commitment aligns with contemporary regulatory standards, ensuring a secure and ethical approach to data handling throughout the research process.

C. *Insights from Performance Benchmarks*

In addition to exploring various aspects of WeStat, the paper meticulously examines its performance through detailed benchmarks. On PC, the third party's ciphers generation exhibits expected linearity with the number of users, while user pre-computation and encryption times are within anticipated ranges. The mobile platform, however, displays an intriguing trend with a smaller on-mobile ratio between pre-computation and encryption times, especially notable with higher user counts. This phenomenon is attributed to potential garbage collector triggers during mobile encryption for larger user groups.

The paper suggests potential optimization by implementing the cryptographic scheme entirely in a native language. Communication costs are also addressed, detailing the sizes of public keys and ciphertexts during registration and participation for different participant counts. These performance considerations contribute to the comprehensive evaluation of WeStat's viability and efficiency in real-world scenarios.

II. IMPLEMENTATION

1) **Set-up**: The Aggregator initializes the system and sets up a binary tree structure with leaves representing users and nodes representing user groups. Cryptographic parameters are generated.
2) **Registration**: Each user registers to the system and is assigned a leaf in the tree structure. They generate public and private keys for tree nodes they are part of.
3) **New Study**: A Third Party initiates a study with the Aggregator. It inserts random values into all tree nodes to blind intermediate results. Study parameters are published.
4) **Data Collection**: On phone, usage data is obtained from apps with user consent. User encrypts own data vector using a functional encryption scheme with keys of tree nodes it belongs to.
5) **Sending Data**: Encrypted data vectors are sent from the user's phone to the Aggregator independently without coordinating with other users.
6) **Aggregation**: The Aggregator identifies a set of tree nodes covering participating users. It aggregates encrypted data at selected nodes, unlocking blinded intermediate sums.
7) **Result Decryption**: The summed encrypted result is sent to the Third Party which removes blinding factors it inserted earlier and decrypts final statistics.
8) **Output Result**: Third Party outputs aggregated usage statistics results to analysts without compromising individual privacy.

III. STRENGTHS OF THE PAPER

A. *Privacy Guarantees through Encryption*

The system uses encryption techniques so that the individual usage data is protected and not revealed to the Aggregator or Third Parties. The encryption scheme ensures that it is infeasible to decrypt any single user's data. It also prevents the Aggregator from computing statistics on data from a single user, ensuring privacy. The use of encryption techniques to protect individual usage data aligns with the paper's goal to 'prevent any inference or re-identification risks'.

B. *Fault-Tolerant System Design*

The paper introduces a fault-tolerant system. Even if some users fail to send their data or decide to opt out after initially agreeing, the system can still function and compute the final statistics over the remaining set of users. This is achieved by using a tree-based structure and only needing to cover the participating users with "target nodes". The dropout of some users does not reveal any partial information.

C. *Balancing Privacy and Useful Statistics*

The solution preserves privacy while enabling the computation of valuable statistics. Despite the encryption of usage data, the system allows for statistical operations such as sums, means, regression, etc., on aggregated encrypted data. This ensures a delicate balance between privacy protection and the capability to derive meaningful insights into mobile usage statistics. Moreover, it guarantees that only authorized aggregations can be performed.

D. *Decentralized Architecture with Minimal Interactivity*

The system adopts a decentralized approach, minimizing the need for user coordination. Each user sends encrypted data independently to the Aggregator. The Aggregator and Third Party interact minimally to set up the study and obtain final outputs, reducing communication overhead.

IV. LIMITATIONS

The paper does not sufficiently delve into the details of the following two limitations:

A. *Data Poisoning Attacks*

Users may voluntarily provide wrong or inaccurate inputs to a study in order to affect its outcome.

This kind of behavior is quite common nowadays in the context of data collection systems. An example is the recently discovered Nightshade tool, used by artists to slightly modify certain pixels in their works in a way that is invisible to the human eye but that makes a big difference to AI models. Its purpose is to counter the use of web scrapers, which collect online images for training of generative tools such as DALL-E or Stable Diffusion [2].

Similarly, given how Third Parties are only able to access aggregate analysis results instead of raw data, participants may misuse specific apps in order to affect such results. These actions would be further motivated by the presence of rewards for contributions.

As a consequence, operations such as the mean (which the paper focuses on) would not find many real-world applications and the WeStat system needs to be able to scale well to more complex ones.

B. *Scope*

The possibility of performing analysis over encrypted data is certainly intriguing. However, being rather limiting for Third Parties, the WeStat system won't become a standard for big platforms anytime soon unless specific laws come into place.

As an example, Meta recently introduced a way for users in Europe to enjoy a tracking-free experience in exchange for a subscription [3]. Given its high cost, almost 13 euros a month via mobile, it is likely that many members will choose to allow data collection to avoid either having to pay for Facebook and Instagram or losing access to them. Consequently, these platforms are already able to perform user tracking without an initial encryption. The WeStat system may coexist with them and be employed on the side, but it seems unlikely that it will substitute current data collection mechanisms in the short term.

C. *System Exploitation*

Moreover, in section 2.2.2 (Security), a model is mentioned in which an adversary could potentially exploit the system by compromising both the Aggregator and the Third Party. This scenario is listed but not considered in the following paragraphs and, consequently, it seems to be left as an unsolved vulnerability.

V. POTENTIAL FUTURE WORKS

A. *Advanced Cryptographic Techniques*

1) **Better Functional Encryption for Data Analysis**: Try to implement homomorphic encryption [4], enabling complex data operations such as aggregations and machine learning on encrypted data. This method will let WeStat conduct in-depth data analytics while maintaining the integrity of user privacy and user safety.
2) **Quantum Computing**: Computers/Processors are getting smarter and more powerful. Try to keep post quantum cryptographic algorithms [5] in mind, including lattice-based cryptography/Multivariate cryptography, to safeguard against potential quantum computing threats, enabling the long-term security of WeStat's data.

B. *Using Artificial intelligence and Data Analytics*

1) **AI in Encryption Management**: Implement machine learning to better handle encryption and decryption [6]. The idea is to let AI check and adjust how data is processed and pipelines based on how sensitive it is and how it's used. By doing this, the system can work more efficiently without giving up on keeping data safe and private. We can use the keywords method, where some words are locked and checked for. The key is to keep the AI learning up-to-date so it always knows the best way to protect data. We can keep the models in limits and set up boundaries on what it can change or access or what it can not.
2) **AI for Spotting Privacy Risks**: Use AI to build models that can spot and stop privacy risks or chances of data breaches before they happen. Setting up authentication channels and disaster recovery mechanisms is extremely important. These models would look at how users interact with the system, how they are sending/receiving and how data moves around, so they can suggest updates or changes to keep everything secure. This means the system can get better at protecting data over time and learn from mistakes.

C. *Staying Current with Data Laws and Regulations*

1) **Adjusting to Legal Changes**: Prepare a team in WeStat which is about legal rules and regulations. They would keep an eye on new data laws like GDPR and CCPA and make sure WeStat follows them. They should use automated systems to change how data is handled whenever there's a new law or rule, especially considering where users are located. Furthermore, keep monitoring the breaches and their post affects, keep updating policies and documents.
2) **Blockchain for Trust and Compliance**: Implement blockchain [7] technology to maintain a transparent record of data usage and processing within WeStat. We could use tools like Non-Fungible Tokens (NFTs) and smart contracts to enhance data tracking and management. This implementation demonstrates compliance with prevailing data protection regulations but also boosts user confidence in the security and appropriate handling of their data. Additionally, it ensures that WeStat's systems align with current industry standards

D. *Making the User First with New Technology*

1) **User Control with Blockchain**: Work on a system where users can manage their own data identity using blockchain [8], access controls and levels will be useful, the tree-like structure mentioned can be implemented. This would give

users more control over their data and make sure WeStat stays focused on keeping user information private and secure.

2) **Clear Understanding with AR**: Develop an Augmented Reality (AR) or a simulation feature in the WeStat app. This would show users a visual or blueprint of how their data is being stored, used and kept secure. By using AR or any relevant technologies, WeStat can make it easier for users to understand the technical parts of data encryption and processing, bringing confidence to them that their data is being handled safely.

VI. CONCLUSION

WeStat's system architecture, implemented by a binary tree of cryptographic keys, supports scalability and fault tolerance, accommodating a huge number of users while maintaining the integrity of data processing and analysis. However, the system is not without its challenges. The complexity of its cryptographic algorithms and methods may create difficulties in broader implementation and user accessibility. Additionally, while it suits well with current data privacy laws, never ending legal and technological changes will require continuous adaptation and improvement to stay relevant and inline with the requirements.

Integrating AI for data analytics and the potential application of blockchain technology for enhanced trust and transparency will be a good path for innovation to follow given the technical evolution. These features, balanced with a user-focused approach, position WeStat not just as another solution for data privacy but as a model that could redefine norms in user data handling and trust in the digital data landscape.

REFERENCES

[1] https://commission.europa.eu/law/law-topic/data-protection_en

[2] https://www.theverge.com/2023/10/25/23931592/generative-ai-art-poison-midjourney

[3] https://www.forbes.com/sites/danielnewman/2023/11/10/meta-launches-ad-free-paid-model-in-europe-does-this-solve-the-privacy-issue/

[4] https://inria.hal.science/hal-02947359/file/2020-197.pdf

[5] https://en.wikipedia.org/wiki/Post-quantum_cryptography

[6] https://link.springer.com/chapter/10.1007/978-3-030-61527-7_27

[7] https://grcoutlook.com/blockchain-as-a-framework-for-trust/

[8] https://www.foley.com/insights/publications/2023/03/what-expect-2023-trends-cybersecurity-data-privacy/