Federal Register Notice 87 FR 35250, https://www.federalregister.gov/documents/2022/06/09/2022-12432/request-for-information-on-advancing-privacy-enhancing-technologies, June 9, 2022

Request for Information (RFI) on

Advancing Privacy Enhancing Technologies

Karlsgate Inc

Karlsgate

Comments on Advancing Privacy-Enhancing Technologies

Respondent	Karlsgate Inc. 3 Kobert Ave Budd Lake, NJ 07828 https://karlsgate.com
Respondent Type	Industry
Respondent Contact	Brian Mullin Founder & CEO

Responding to topic 2. (*Specific technical aspects or limitations of PETs)*

Privacy Enhancing Technologies (PETs) can become a great aid to responsible data management in both the public and private realms. Privacy and the control over who becomes aware of personal identity has protective, ethical, and legal implications.

NEW EMERGING TECHNOLOGIES

An emerging technique called Partitioned Knowledge Orchestration, which is in a family of secure multi-party computation technologies. This is particularly useful for data sharing operations where the ***disassociation and confidentiality of identity*** is critical. The defining feature of this approach is the deliberate fragmentation of any complete information that could lead to either identification or re-identification. This partitioning is created during a careful orchestration of transactions performed by at least three independent actors. Of the three actors, two will be data controllers that compute identity hashes, and one will be a blind facilitator, which receives identity hashes from each data controller for comparison but never receives the one-time-use formula used to build the hashes. This methodology is particularly useful when considering how to protect the data from re-identification risks.

THE HEART OF THE PROBLEM: DATA LINKAGE

Partitioned Knowledge Orchestration was purpose-built as a solution for the critical **linkage problem**, which can be defined as:

> Given, two independent entities (public or private) are each managing a dataset about individuals. The understanding of each individual's identity is achieved using various identifiers such as name, postal address, email, and/or social security number. However, these components of personal data are sensitive and are tied to personal privacy rights, regulatory restrictions, and/or ethical handling concerns.
>
> How to enable the 2 independent entities to share the understanding of the individuals in common between the 2 datasets without sharing any personal data and without inadvertently allowing reidentification of those individuals not in common (i.e., outside of the desired intersection)?

Some real-world examples of the linkage problem (where data privacy and data sharing are equally critical) include:

- Retrieving Protected Health Information (PHI) associated with a patient in the form of Electronic Health Records (EHR) from another health system.
- Detecting fraud or anti-money laundering activity between banking institutions
- Researching rare disease treatments by gathering longitudinal views of patient data
- Detecting duplication in voter registration databases
- Anonymizing contact tracing interactions for viral exposure monitoring

SHARING DATA SAFELY & SECURELY

The linkage problem causes many challenges when it comes to data collaboration. Whenever records with individuals as the data subject are joined together, solutions commonly used today require that one party needs to fully trust the identity of subjects with the other party. Sometimes that is due to direct, clear text data sharing. Other times, pseudonymization techniques are used. It is important to note that pseudonymization enables reidentification when combined with additional information and is defined clearly as "personal data" under GPDR regulations in the European Union. Several categories of PETs protect the privacy of data but still serve as a form of pseudonymized data when treating identifying data.

Another consideration for advancing PETs is the need for prolonged public and government sector evaluation of the safety, fidelity, and reliability of new cryptographic technologies. Since Partitioned Knowledge Orchestration reuses proven cryptography in an orchestrated manner, it has a key advantage in institutional settings. There is no need to review, analyze, and vet new and emerging cryptographic algorithms, each with its own weakness profile. Instead, the sequenced orchestration can rely completely on proven, FIPS-compliant cryptography that can be substituted with other approved algorithms over time in a very natural evolution. Again, **this is not a 'black box' technique** but an orchestrated framework of interactions with defined actions including secret key derivation, one-way hashing, and symmetric encryption transformations.

EMPLOYING FIPS COMPLIANT CRYPTOGRAPHY TO SOLVE THE LINKAGE PROBLEM

Some common choices for the cryptographic algorithms are: Elliptic Curve Diffie–Hellman (secp384r1) for the secret key derivation covered by FIPS PUB 186-4, HMAC-SHA-384 for the one-way hashing covered by FIPS PUB 180-4, and AES-256-CBC for the symmetric encryption covered by FIPS PUB 140-2. **The use of well-studied and FIPS-compliant cryptography can accelerate adoption and application of PETs in government settings.**

Several observable attributes of Partitioned Knowledge Orchestration can satisfy the design goals of data sharing operations that are vulnerable to the linkage problem.

- Zero-trust required of exchange partners including actors in facilitating roles, since no other party ever receives information that can lead to re-identification. This forms a stronger information security stance that cannot be compromised due to never enabling exposure. There are no data artefacts or exhaust from an exchange that can be retained to leverage for re-identification attacks.
- Zero change in data custody preceding, during, or subsequent to the sharing transaction. This is vital to establishing provenance and permissioning over data assets without back-door allowances for data sharing operations.
- Zero obfuscation of the collaboration methodology including cryptography, matching rules, or expected outcomes. This leads to faster adoption of the PETs and higher levels of trust among government entities, private industry, and the public at large. New cryptography poses higher risk of undiscovered vulnerabilities. A transparent approach is superior to an opaque approach when cryptanalysis and wide-spread acceptance is required.
- Zero limitations on topology of the exchange network promote an open environment that can cross technological, industrial, and geopolitical boundaries. Since the data intersection architecture is completely distributed and decentralized, there is no limiting factor on capacity. Every transaction enrolls 3 actors: 2 encoding nodes and 1 comparison node between them. Each of these actors can be commissioned and decommissioned at will including once per transaction. This approach is a natural fit for global scale and open architecture.

COMPARING THE DATA PROTECTION METHODOLOGIES

The following table details the comparison of common PETs for personal identity linkage applications:

Encryption Using a secret key, information is scrambled until the key is re-applied	• Only protects data in transit • The recipient will decrypt back to fully identified data upon processing • Full custody change
Hashing / Tokenization One-way scrambling of data that is very difficult to reverse	• Full change of custody leads to uncontrolled, future re-identification attempts against an identity graph • Data owner has no agency over copies of pseudonymous data and that is why the practice is not GDPR-compliant

Data Clean Room Full dataset is sent to a controlled and isolated environment that is typically managed by a third party	• Each usage represents a full custody change event • Both parties must trust the same operator with full vetting, legally binding responsibilities, consent and security obligations • Difficult to have a single clean room service for all partners and there is no clean room-to-clean room exchange mechanism
Fully Homomorphic Encryption An advanced form of cryptography that allows analysis of data without decrypting the payload	• Does not solve the linkage problem; while the exchanged attributes are protected, the payload can still be attacked for re-identification through joining to a known identity graph • Performance problems make this technique max-out at ~5 million records practically speaking
Federated Learning Building an aggregated model constructed from multiple, localized machine learning processes	• Does not solve the linkage problem; while powerful in building audiences and propensities in a privacy sensitive manner, it does not produce a deterministic intersection between two identified data sets, therefore limiting its use cases
Differential Privacy A process of adding randomly generated adjustments to data without introducing significant skew	• Does not solve the linkage problem; while it can be helpful for obfuscating descriptive values, it has no purpose for directly linking one identity to another • May degrade accuracy for certain use cases
Partitioned Knowledge Orchestration A coordination to share common identities while simultaneously blocking unwanted reidentification as a consequence of interacting	• No identifying information, including ciphertext or hashes, ever flow directly from partner to partner • The facilitator is deprived of any cryptographic parameters, blinding it to any ability beyond simple comparisons • No trust is required to safely connect data sets and extract insights, since no identity information is put at risk

A NOVEL APPROACH DESIGNED BY KARLSGATE

Karlsgate has developed a practical implementation of the Partitioned Knowledge Orchestration technique called Triple-blind Identity Mapping Protocol (T-BLIMP), specialized for linking identity information at scale.

The orchestration procedure of T-BLIMP follows a simple sequence.

1. The orchestration of events starts with managing single-use cryptographic keys controlled solely by the trading partners.
    a. Each party randomly picks a seed value (a very large number).
    b. A shared secret is derived between the 2 partners via the Diffie-Hellman algorithm and is never shared with anyone else.
    c. Identifier values are extracted from the raw data, e.g., email address.
    d. Single use Hash Tokens are assembled from the identifier values + the seed values + shared secret + the selected cryptographic hashing algorithm. (Hashed Tokens are different for every trade.)
2. The next step enlists a third party, the facilitator, to stand between the trading partners to perform neutral and naïve adjudication of hash collisions.
    a. Only Hashed Tokens are transmitted to the Facilitator by each trading partner.
    b. The Facilitator compares the lists of Hashed Tokens and sends back a signal on each matching value directly to the trading partners. (The Facilitator never knows the math used to compute Hashed Tokens.)

This technology is fully developed, commercially available, validated by industry, and in international use at scale.

Karlsgate has been commercially offering privacy-enhancing dataset intersection services, based on T-BLIMP, since early 2020. The enterprise-scale solution is already processing data sharing transactions among very large data sets with over 1 billion identity comparisons per transaction. Usage of the service spans the globe including users in United States, Mexico, and Australia. The most common industries using the solution to date include healthcare, financial services, and digital marketing.

PRIVACY ENHANCING TECHOLOGY FOR THE PUBLIC SECTOR

The use of PETs can have a major positive impact on the safe and ethical use of governmental and public institutional data. However, as described above, many of the methodologies that are commonly used today have made limited, if any, progress with respect to solving the linkage problem inherent in majority of the use cases where PETs could be of value. Partitioned Knowledge Orchestration is the ideal technology for meeting the public sector need. It enables the linking of identifying data without ever compromising privacy or security.

Responding to topic 3. (*Specific sectors, applications, or types of analysis that would particularly benefit from the adoption of PETs)*

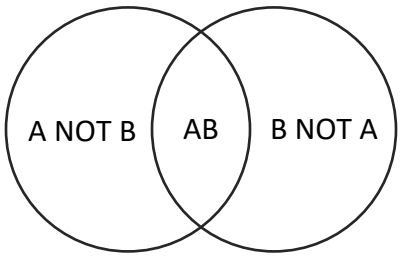
Data collaboration use cases continue to grow across all sectors and verticals as the value of data is unlocked with ever evolving technologies that enable leveraging data for data sharing, behavioral tracing/tracking, artificial intelligence, analytics, planning, reporting, targeting, measurement, personalization, etc. With data collaboration comes the need to be able to **link data without enabling any form of re-identification**.

With a challenge that is so pervasive, it would be unwieldy to create a comprehensive list of all potential applications that would benefit from adoption of PETs. However, the following applications illustrate a few of the types of applications that would particularly benefit from the adoption of PETs which specifically solve for the **linkage problem**:

> *NB: Special care is needed to avoid technologies that include or depend on pseudonymous keys. Many PETs only obfuscate identity through the use of pseudonyms (i.e., directly passing data values between partners that serve to uniquely reference individuals). The possession of pseudonymous keys inherently enables re-identification immediately or anytime in the future.*

- Healthcare HIPAA compliance when sharing patient data while eliminating the transfer of specific Protected Health Information
- Cross-border data exchange where jurisdictions apply differing regulations on the Right to Privacy and use of Personal Data
- Financial fraud blacklisting and other domestic and international sanction lists
- Screening traveler identities against the FBI Terrorist Screening Center No Fly List
- Rare disease treatment research using Real-World Data (RWD), which can support and/or complement data collected from traditional clinical research
- Contact tracing operations for public health purposes
- Checking against FDA Debarment lists or other similar regulatory watchlists

A common requirement in many of the above use cases is that 2 organizations (with data sets ***A*** and ***B*** respectively) need to find the identities belonging to set ***AB***, while never revealing the identities belonging exclusively to either set ***A NOT B*** or set ***B NOT A***.



References:

- **GDPR** - EU General Data Protection Regulation https://gdpr-info.eu/
- **FIPS** - NIST Computer Security Resource Center https://csrc.nist.rip/publications/fips
- **T-BLIMP** - Karlsgate Identity Exchange https://karlsgate.com