GEORGIA POLICY LABS

Georgia State University | Andrew Young School of Policy Studies

Using Common Hash Values as Linking Keys

A Solution for Identifying Linkage Keys (SILK) Use Case

Robert McMillan and Chris Thayer

November 2021

Acknowledgment

The project described herein was supported by the Human Services Interoperability Innovations program, grant number 90PD0308, with the amount of $599,992 from the Office of Planning, Research and Evaluation, Administration for Children and Families, U.S. Department of Health and Human Services.

Its contents are solely the responsibility of the authors and do not necessarily represent the official views of the Office of Planning, Research and Evaluation, the Administration for Children and Families, or the U.S. Department of Health and Human Services.

Executive Summary

This document describes an illustrative yet hypothetical use case for SILK (a Solution for Identifying Linkage Keys), which is a tool that helps agencies securely link their data without sharing sensitive fields. The document provides

- a description of current data sharing challenges;
- an introduction of the SILK solution;
- the linking process using hashed values;
- a discussion of the challenges, risks, and management requirements of SILK; and
- an example use case scenario.

The core data sharing challenges addressed by SILK are the security risks associated with housing personally identifiable information (PII) and obtaining legal approval for PII sharing between partner agencies. Sharing sensitive identifiers like social security numbers (SSNs) or other private identifiers are crucial for accurate linkage between datasets; however, holders of data with such fields are often subject to intense legal and privacy constraints that prevent sharing SSNs (or other identifiable data elements) with partner agencies.

SILK solves these challenges by transforming an SSN or other sensitive value into a secure hashed value using a secret key called a Salt. Any SSN hashed with the same Salt will always produce the same hashed output. Two (or more) partners can independently use SILK with the same Salt to convert their sensitive fields into secure hashed values. The underlying data remains secure because the Salt value is kept secret, and it is too large and random to be guessed in a meaningful amount of time, even with supercomputers.

The benefit is that the partners can now link their datasets using the hashed field combined with other less-sensitive non-identifiable data while removing the need to share SSNs to link records. This eliminates the security risks and legal challenges because no private information is shared.

This document includes a more detailed discussion of the challenges, risks, and management requirements associated with using SILK. The described challenges and risks are largely the same as those involved in using SSNs to link datasets. This document also provides recommendations for the mitigation of these

challenges and risks. SILK's management needs are minimal, limited to steps needed to set up and maintain the tool for use.

Finally, the example use case scenario demonstrates the benefits of SILK by describing a situation where two partner agencies—a state department of education (DOE) and a state social safety net benefits agency (DOSS)—need to link their data to issue benefits. With SILK, neither agency has to share the SSNs in their datasets. Instead, sharing data with hashed values allows linking in the same manner as SSNs but without the risks or legal limitations. SILK makes the process of agency partnership faster, easier, and safer by reducing the sensitivity of the data the two partners provide, thus letting them link their datasets and issue benefits faster and more efficiently.

Introduction

Human service agencies serve people who interact with various state entities. Each agency's administrative data (i.e., data collected on individuals to facilitate service provision), however, are regularly housed separately, preventing coordinated decision making and resource allocation. As human service agencies strive to implement evidence-based programs and policies, they need a more holistic view of their constituents' lives and their data, which requires interoperability. Interoperable data systems are those that can "talk" to each other, achieved by using unique per-person identifiers to allow agencies to connect disparate data. By connecting these separate data pools, service providers can understand better the full range of systems with which their clients are interacting and more holistically assess current service use, risk factors for program failure, opportunities for improved efficiency, and evolving client needs.

Elements of Partnership

To achieve such efficiencies, interested agencies must undergo a partnership process. This process requires building relationships among the potential participant agencies, establishing trust, and identifying opportunities for mutual benefit. Once the agencies identify an opportunity for collaboration, they must determine what data they will need to share with one another to achieve their goals. This includes program-specific data and the identifying fields necessary to

link two or more dataset that do not have a single administrative identifier in common.

After the agencies identify the necessary data and its purposes, they can begin drafting a data sharing agreement (DSA). A DSA is a document that outlines the relationship the agencies have built, describes the contents of the data they have decided to share, and delineates requirements around the protection and privacy of that data (e.g., transfer methods, storage needs, deletion protocols, and other technical requirements to allow the safe handling of any shared data). After the DSA is in place, the partner agency or agencies can transfer data to whichever entity will be doing the data processing. The datasets can then be linked to one another, and analyses performed on the combined datasets.

SILK eases this partnership process in two key ways. First, it reduces the amount of sensitive data a partner needs to share by turning a sensitive PII value (like an SSN) into a secure hashed equivalent. Other PII, such as names and/or birth dates, will still need to be shared to enable matching. Second, SILK technology makes it easier to establish a DSA and the relationship it embodies because the amount of PII and its degree of sensitivity can be reduced through SILK. The technology also mitigates barriers to partnership by reducing the risks involved, thus facilitating trust.

Connecting Person-Level Data from Multiple Agencies

Each of these organizations' datasets are governed by separate confidentiality rules regarding the use and transmission of PII, a required component for effective matching of records for the purpose of linking datasets used to accurately calculate household benefit amounts. Before organizations can share data between one another, they must complete DSAs outlining the specific data elements and their terms of use. The process of negotiating the specifics of DSAs is notoriously lengthy and complex, due in part to concerns about the sensitivity of data elements to be transmitted.

SILK helps dissolve these concerns by providing an alternative to sharing highly-scrutinized PII variables like SSNs.

Interoperability and Sensitive Indicators

After partner agencies establish trust, they frequently encounter a significant barrier to achieving system interoperability and the benefits it can provide. In many internal agency data systems (particularly for long-running or "legacy"

systems), the primary identifier for each person is their SSN. SSNs are an important part of the PII that documents who they are and distinguishes them from other people in an agency's records. While the SSN is a high-reliability identifier within a data pool, it is also heavily affected by legal privacy constraints. Regarded as the most sensitive indicator of an individual's identity, agencies do not typically or easily share data that contain SSNs.

Many agency datasets do not have other (reliable) values in common. Agency-generated identifying codes are naturally agency-specific, while commonplace identifiers (e.g., names) may apply to multiple people and might change over the course of an individual's life. This means the SSN would be a powerful tool for successful linking if it were shareable. This barrier to data sharing halts DSA development, weakens budding agency relationships, and perpetuates siloed decision-making. It also prevents human service agencies from more effectively understanding multi-agency client experiences, analyzing problems, and implementing evidence-based solutions.

SILK technology removes the SSN interoperability barrier, thereby facilitating linkable datasets for research and practice purposes without needing to release select key PII values like the SSN.

Breaking the Interoperability Barrier

The Administration for Children and Families grant award funded the proof-of-concept (POC) implementation of a hashed SSN to be used as a proxy ID. The POC consists of a client Python application designed to convert single records or batch files transferred to a cloud environment (Microsoft Azure) for processing and long-term storage. SILK is evolving into a system-to-system Application Programing Interface (API)-based solution proven to automate an identity-matching solution that human service agencies can use to eliminate sharing SSNs, allowing for the easy, secure matching of individuals across agencies and the sharing of de-identified data with research partners. Its features include a file-based search function that can perform individual or batch searches and can return an unlimited number of matching records. At this time, the search function cannot provide counts or summary statistics, but as an open-source solution, additional features could be added as needed.

This hashing solution gives a data provider a method of hashing any specified variable into an irreversible identifier that they can then use as a linking key to other datasets. The variable content itself (the SSN or other secure value) is

never transmitted to the receiving party or stored in the central database. This open-source solution can be used by an unlimited number of data providers for the purpose of linking datasets across organizations for research and analysis. Any number of partnering organizations may use the solution if at least one has the Azure subscription needed to host the instance. Any organization using the solution does not need to involve the Georgia Policy Labs (GPL) for it to work. It is important to note that a common Salt is needed for matching. Different instances with different Salt values will not produce linkable datasets.

Using an API-based Proxy ID Generator for Linkable Datasets

The intent of the solution is to provide a cloud-based platform for various agencies in the health and human services category and other domains to share non-PII data with other agencies securely. The solution provides the ability to identify and connect all the information that is available between the various agencies uniquely using important PII identifiers (such as SSN) without having to share those identifiers. These are specific unique identifiers corresponding to the non-PII administrative data that needs to be shared and linked to facilitate a larger inter-agency task, such as resolving and issuing Pandemic Electronic Benefit Transfer (P-EBT) benefits as outlined in the use case. Using common hash values, the agencies can engage in a linkage process (next section) that accurately connects their datasets while keeping PII unshared.

The common hash values creation tool—SILK—serves this function. It uses API technology to connect to a secure server that hashes data. An API solution allows for systems to securely generate alternative, unique, and linkable hash values derived from the original ID. The source system (the agency) authenticates to the API (i.e., logs in with their secure credentials) and submits the PII variable. The secure server automatically processes that ID and generates a hash value—a 150-digit code—based on the Salt selected (a secret key value). A different organization with the same PII value, processed through the same SILK software but using a different Salt, would receive an entirely different hash value. This means that agencies can be confident that only their trusted partners, using the same pre-agreed-upon Salt keys, will be able to connect hash values and, therefore, records.

The agency never has to worry about interacting with complex programs or determining Salts directly. They simply point the SILK API at the desired file, indicate their key, and receive a hashed value for each PII value, which they then associate with their original data record(s)—replacing the sensitive value with its hashed alternative. The full record can then be shared in a central repository or

among trusted data partners. This allows data to be connected across multiple agencies that are using common hashed identifiers based on the same Salt key to create a more holistic picture of all the master data collected around an individual or single entity without having to share sensitive PII with any other organization.

Linking Process Using Hashed Values

Within a single agency's data pool, individuals are typically identified by a single, authoritative code in every dataset (e.g., table) in which they appear. While this code is fully reliable within the agency, other agencies do not share it and have their own codes to identify records in their systems. When two datasets lack a common code, the datasets must instead be linked based on the PII of the individuals observed. When the datasets have many PII fields in common, it is easier to be confident in the accuracy of a match. Fewer fields, fields that are not well-populated (i.e., many blank entries), and low-reliability fields (i.e., fields with a high risk of a typographic or other error) make matching more difficult.

The most common linking criteria are first name, last name, date of birth (DOB), and SSN. On their own, the first three criteria are relatively weak identifiers. Name values are highly susceptible to punctuation and spelling variations, common first-last sets may be shared by many people, and names may substantively change over time due to marriage, adoption, divorce, etc. Likewise, dates of birth are vulnerable to typographic errors and shared by many people. However, when these weak identifiers are combined—especially with the inclusion of a high-quality identifying value like SSN—they can produce very reliable matches. Indeed, SSN is such a sufficiently high-quality field that it can be combined with only two of the other common criteria and still produce a confident match, increasing the likelihood of detecting and appropriately matching records where, for instance, a name has changed, a DOB digit was incorrectly entered, or some other barrier prevents a match based on first name, last name, and date of birth. By hashing SSN values, partner agencies can enjoy a simplified matching process and increased match reliability of linking with a high-reliability field without having to share the SSNs.

Remaining Challenges of Probabilistic or Hand Matching

Of course, the use of hashed SSN values does not remove all challenges with matching approaches—whether executed using probabilistic techniques or "by hand." The best practice for matching is to seek a full match on every common

field between the two datasets. As mentioned previously, though, there will virtually always be some proportion of records that will not match on all fields present. This is the point at which matching expertise must take over. That expertise might be in establishing the criteria and algorithms for probabilistic matching routines, which compare unmatched records and calculate their degree of likely match as measured by metrics such as number of characters in common in each match field. Alternatively (or in parallel), match expertise might be exercised by individually inspecting (i.e., hand matching) the remaining unmatched records and considering partial matches, such as on DOB, SSN, and first (but not last) name, to identify valid matches blocked by a typographical error or unrecorded name change.

Risks and Management Requirements

Risks/Potential Errors

While the SILK system makes the hashing process itself reliable and secure, there are still several sources of potential error of which to be aware when matching using hash values. Two sources originate in any use of SSNs to link, while only the third is SILK-specific.

One challenge of hashed SSNs (also present when using traditional SSNs) is substitution. It is not uncommon for parents of young children who may not yet have their own SSN values to instead provide the parent SSN when submitting data for the child. When processed through SILK, the parent's and child's records will consequently produce the same hash value, and this means an SSN-only match would erroneously pair datasets across generation. However, as with working with SSNs directly, this error can be avoided by relying on additional fields (particularly DOB) to disentangle such records.

Second, there is the matter of "code" SSNs. Many agencies may use SSN values that are, in fact, codes (e.g., 000000000, 121212121, 616161616). These codes indicate to the agency using them something about the record, such as a supposed teacher entry (on a student's record) being a substitute or institution, a supposed head-of-household lacking a valid SSN, etc. While such codes are fairly obvious when working with raw SSNs, such codes will hash to values that look equally valid to the hashes of true SSNs. To avoid this difficulty, it is important for partner agencies to provide one another with a list of any code SSNs. The agencies can then run these lists through the standard SILK process

to establish a list of code hashes that their matching approach can identify and appropriately flag as code.

Finally, there is one point of risk specific to SILK-processed SSNs: the consequences of typographical errors within the SSN value. Like any other field, errors are possible any time a human is required to enter data. If the SSN value in one partner's data has a typographical error, the resulting hashed value will be very different from what the hashed value would have been if the SSN were correct. This means that the two hash values will not match up when the datasets are compared. Unlike when working with raw SSNs, it is not possible for a probabilistic evaluation or individual inspection to identify an obvious typographical error, so the records will remain unmatched if their other match field values are insufficient to establish confidence in the potential match.

Necessarily, these risks or caution points only apply when SSNs are the hashed value. If hashing another identifier, the particulars may change somewhat (though the typographic and code value points of care may be similar).

Management Requirements

After the initial setup, operational management requirements for the SILK solution are quite low given it is open-source software. The participation of multiple agencies requires DSAs detailing which datasets from which partners are allowed to use programmatic data from each other's datasets. Each agency uploads its datasets to the central datastore. Because the hashing solution is centralized, the agencies will be able to link their data together using a common hash value without ever interacting directly with the protected sensitive identifier or other PII. The DSA governs access to the joint data results, which is configured through the search function of the solution.

Management Considerations

1. Operations management (to whomever this role is assigned) establishes an Azure-hosting instance of SILK. This use case describes an Azure subscription hosted by GPL, but the solution could be hosted by any group wanting to partner in the deployment of the solution. While SILK is free and open source, whoever is hosting the Azure instance must pay the Azure subscription cost. This may be a shared cost between data providers, data consumers, or whatever arrangement is negotiated. The cost can range anywhere from a few hundred dollars a month for a minimum installation to a few thousand for a multi-agency engagement.

- An intermediate to advanced level of Azure knowledge is needed for configuration, including
    - Azure SQL Server,
    - Azure virtual machines,
    - Azure security,
    - Azure Active Directory,
    - Azure backups, and
    - deploying Azure solutions.

2. Issuance and maintenance of secure credentials to access the Azure instance hosting SILK
    - Once acquired, the only maintenance necessary would be updating them if/when transitioning SILK duties to a new employee.

3. Selection and documentation of which Salt key the agency and its partner(s) will use in SILK

Secure transmission and storage protocols are TLS 1.2 and FIPS 140-2 compliant. Anything needed beyond those already in place should be specified early in the DSA-drafting process.

Scenario: Connecting Agencies to Calculate P-EBT Benefits

This use case explores a situation requiring cross-agency dataset linking: the provision of P-EBT funds to all eligible children in a state. Due to the COVID-19 pandemic, many schools and child care facilities were closed or had reduced attendance capacity or hours. During normal operation, these facilities would provide eligible students nutritional assistance. P-EBT benefits are designed to replace the meals missed due to pandemic-related closures. In some states, P-EBT benefits had to be issued with multi-agency cooperation that lacked existing data sharing agreements or unified record keeping systems. In this use case, DOSS, the issuing agency for SNAP benefits, does not hold any data about the population of school children in the state (that are not part of SNAP households). DOSS must partner with another agency, the DOE, to identify

school children eligible for P-EBT via SNAP and free or reduced-price (FRM) eligibility and gather the data necessary to issue these benefits.

Uniting Two Sources of Eligibility

To issue P-EBT benefits, DOSS requires a household-level file containing the total benefit amount to be issued for current SNAP clients and an individual-level file for non-SNAP eligible students to be issued new EBT cards. The benefit is calculated based on 1) the number of eligible children in the household and 2) how much (the percentage) of each month of facility-provided meals each child missed due to pandemic-related closures. Although DOE has the FRM-eligible schoolchildren, those data must be matched with SNAP clients from DOSS prior to benefits administration. The two data pools—each held by a different agency and with its own identifying codes—must be united:

1. the population of children in SNAP-eligible households (hereafter "SNAP," provided by DOSS) and
2. the population of all children in a public school in the state (hereafter "DOE," for its originating institution) who are FRM-eligible.

Establishing single-population eligibility is relatively simple, as it relies on each agency connecting its own data to non-sensitive geographic information on facility closures. The DOE data already have school district information on each student, meaning it can simply be united with a list of school open statuses by month to yield eligibility data. DOSS data already has county of residence information on each SNAP recipient, meaning it, too, can be easily connected to a geographic dataset on facility statuses. However, the subsequent steps involving individual-level data are much more challenging to connect and require the use of PII to establish identity.

In this use case, SILK was not yet available for implementation. Instead, organizations had to form DSAs with a third-party research organization to house and connect these data. Due to the sensitive nature of the data, these agreements and security protocols caused major delays. This is where the common hash value solution dissolves barriers.

P-EBT Data Processing Overview with SILK

1. DOSS prepares its data by
    - generating SNAP file for all children, complete with head of household

information and full addresses (including county) and

- pairing county-enabled individual SNAP data with county-level facility openness records to calculate P-EBT benefit per SNAP child.

2. DOE prepares its data by

   - receiving person-level data from each district;
   - compiling districts' person-level data into DOE file, complete with full addresses (including school district); and
   - pairing district-enabled individual DOE data with school openness data to calculate P-EBT benefit per schoolchild.

3. DOSS uses SILK to hash its sensitive field, SSN, simply and securely.
4. DOE uses SILK to hash its sensitive field, SSN, simply and securely.
5. DOSS compares its data with DOE via the linking process without either party ever sharing or transmitting SSNs.
6. DOSS removes duplicate entries to prevent over-issuance of benefits.
7. DOSS sends data through its established benefits issuance process, helping feed hungry children.

Benefits of Processing Using SILK

Without SILK to securely hash SSN values, thus obviating the need to share the SSNs, the partner agencies had to

1. spend additional time and effort establishing the deep trust required to share SSNs;
2. establish technical protocols for transferring data with SSNs, requiring additional time and effort;
3. determine safe storage and/or destruction protocols for data with SSNs; and
4. extend additional DSA development time to record the solutions to the foregoing problems.

Even with existing relationships among all agencies, this delayed human service operations. As agencies think about adopting SILK, they should understand it can be useful for expected projects and goals and be of tremendous benefit when unexpected needs or emergencies arise.

Conclusion

As the use case demonstrates, the SILK solution facilitates data sharing between agencies. By converting a highly sensitive PII indicator like an SSN into a securely hashed value, SILK helps reduce the technical and legal strictures that can make data linkage difficult, which lowers the interoperability barrier. In turn, this lower threshold for successful data sharing makes DSA development easier as it reduces the technical constraints on secure data sharing. Easier DSA formulation saves time and effort, promotes positive agency relationships, and supports efforts for interoperability, cooperation, collaboration, and enhanced service provision.

Appendix

Source Code

McMillan, R. & McKee, C. (2020). *INOIS Background Worker* [Github repository]. Public. github.com/rmcmillan4/inois-background-worker

McMillan, R. & McKee, C. (2020). *INOIS Desktop Client* [Github repository]. Public. github.com/rmcmillan4/inois-python-client

Installation Manual

Georgia Policy Labs. (2020). *INOIS - Interoperability Network for Operational Intelligence Solution*. dropbox.com/s/kqofu5o9g74nrmk/IINFOIS-Tech%20Docs%20and%20Installation%20Manual.pdf?dl=0

About the Authors

Robert McMillan

Robert McMillan is the director of data integrity for the Georgia Policy Labs. He leads a team of data scientists who are responsible for curating data for research opportunities. He provides technical guidance for the collection and storage of information from GPL's local and state partners. He has over 25 years of experience with enterprise system design, integrations, and operation. Twenty of these years have been at the director level or as a principal consultant focusing on the delivery of large complex new builds, legacy migrations, and more recently with cloud solutions. He was a senior policy analyst at the Georgia Governor's Office of Planning and Budget. He previously held director-level positions at the State Data and Research Center at Georgia Tech and the Department of Early Care and Learning. As a principal consultant in the private sector, he has led multiple state agencies through modernization efforts and has worked with the Department of Defense to coordinate system-of-systems requirements management for U.S. Army training and simulation solutions.

Chris Thayer

Chris Thayer is a research associate with the Public Finance Research Cluster (PFRC) and the Georgia Policy Labs (GPL), specializing in data preparation, de-identification of sensitive data, and science and policy communications. Their research interests include administrative data linkage and housing programs. Chris received a master's degree in public policy and a master's degree in city and regional planning from the Georgia Institute of Technology and bachelor's degrees in business administration and English from Otterbein University. Prior to joining Georgia State University, they interned at the Federal Reserve Bank of Atlanta and served as graduate research assistant at the Center for Urban Innovation at Georgia Tech.

About the Georgia Policy Labs

The Georgia Policy Labs is an interdisciplinary research center that drives policy and programmatic decisions that lift children, students, and families—especially those experiencing vulnerabilities. We produce evidence and actionable insights to realize the safety, capability, and economic security of *every* child, young adult, and family in Georgia by leveraging the power of data. We work alongside our school district and state agency partners to magnify their research capabilities and focus on their greatest areas of need. Our work reveals how policies and programs can be modified so that every child, student, and family can thrive.

Housed in the Andrew Young School of Policy Studies at Georgia State University, we have three components: the Metro Atlanta Policy Lab for Education (metro-Atlanta K-12 public education), the Child & Family Policy Lab (supporting children, families, and students through a cross-agency approach), and the Career & Technical Education Policy Exchange (a multi-state consortium exploring high-school based career and technical education).

Learn more at gpl.gsu.edu.