DEMONSTRATION OF A PRIVACY-PRESERVING SYSTEM THAT PERFORMS AN UNDUPLICATED ACCOUNTING OF SERVICES ACROSS HOMELESS PROGRAMS

by

Latanya Sweeney, PhD

Associate Professor of Computer Science, Technology and Policy
Director, Data Privacy Lab
School of Computer Science
Carnegie Mellon University, Pittsburgh, PA
latanya@cs.cmu.edu

FINAL REPORT

U.S. Government Release

October 2008

Table of Contents

Index of Figures

1. Executive Summary

Over the last two years, the United States Department of Housing and Urban Development ("HUD") reviewed ways to perform a national unduplicated accounting of visit patterns across homeless programs, while respecting the confidentiality of those clients who visit domestic violence homeless shelters.

The goal of the work reported in this writing was to demonstrate a system that performs an accurate unduplicated accounting across homeless programs with guarantees of privacy protection for clients of domestic violence homeless shelters.

HUD sponsors locally administered Homeless Management Information Systems ("HMIS") in order to collect data needed for an annual report HUD provides to Congress termed the "Annual Homeless Assessment Report (AHAR)" [1]. A HMIS is a computerized data collection and processing system designed to capture person-specific information over time from homeless persons being serviced by any homeless program, including domestic violence homeless shelters. Information gathered from all homeless service programs that are geographically co-located is compiled by a HMIS operated by a "Planning Office" (called a "Continuum of Care" or "CoC" in HUD documents) that is local to those programs. Information collected at homeless programs is not directly forwarded to HUD. Instead, the local Planning Office de-duplicates and forwards de-identified, unduplicated aggregate information to HUD.

Special privacy considerations are given to the clients of domestic violence homeless shelters so that client information provided by a domestic violence homeless shelter to a HMIS cannot be re-identified to the clients who are the subjects of the shared information. HMIS are to gather information from local domestic violence homeless shelters in such a way that client confidentiality is maintained yet an accurate unduplicated accounting of visit patterns can still be achieved across homeless programs by planning offices.

In initial steps to protect privacy, HUD modified the fields of information it recommends domestic violence homeless shelters share with a HMIS [1]. The fields HUD recommends are termed the "Universal Data Elements." Rather than using client names or Social Security numbers in the Universal Data Elements, HUD introduced the notion of assigning a unique identifier ("UID") to clients of domestic violence shelters [2].

This paper reports on the use of a technology ("PrivaMix") for constructing UIDs and performing de-duplication such that an accurate unduplicated accounting results while protecting the privacy of the clients who are the subjects of the UIDs (Section 8).

In a real-time experiment, a "PrivaMix Demonstration System" computed an accurate unduplicated accounting using real-world data from homeless programs in Des Moines, Iowa ("the Iowa Experiment"). This writing examines the experiment (Section 10), the data elements shared, the client information used to construct UIDs , the algorithms that generated those results (Section 9), and the privacy implications of results (Section 11).

Here is a summary of performance findings.

The PrivaMix Demonstration System introduced no errors in the unduplicated accounting. It performed exactly as if plain text was used even though Client information was provably never shared with the Planning Office or the other shelters. Section 9 introduces the PrivaMix Demonstration System. Section 10 reports results from the Iowa Experiment.

The client's combination of {*date of birth*, *first three letters of first name*} were used to generate secure UIDs. This writing terms this the "Privacert encoding" as Privacert first proposed its use. Experiments compared Privacert's proposed method with using Social Security numbers, and two methods currently in use by Servicepoint[1]. Privacert's method encountered fewer fields having omissions or errors than the other methods, and used fields in which clients provided more consistent values than the fields used by the other methods. In performing an unduplicated accounting, the Privacert method proved more accurate than the other approaches. Section 10 reports on a comparison of the use of demographics in forming UIDs. (While Privacert proposed this encoding, it is important to note that the PrivaMix System is not specific to any particular encoding method.)

Modifications to the shared data elements improved privacy without loss of reporting ability. Participants in the Iowa Experiment shared *year of birth* with the Planning Office instead of the full month, day, and year of birth as currently recommended in the Universal Data Elements. Doing so, reduced the likelihood of re-identification using publicly available data from 87% to 0.04% (see Section 4.5). While this is an important improvement, other privacy threats remain in the data elements (see Section 11) and are further discussed below.

PrivaMix guarantees privacy protection for UID creation and use in de-duplicating. As noted above, these privacy protections had no adverse effect on de-duplication. However, privacy threats related to the selection of which client-level data elements to associate with UIDs remains. These problems reside beyond the scope of the PrivaMix Demonstration System (or any other UID technology). Below is a discussion of these vulnerabilities and a description of how post-processing anonymization can be added to the PrivaMix System to remedy them.

Data linkage vulnerabilities exist when a Planning Office subsequently shares de-duplicated results with the HMIS (Section 8). Collusion between the HMIS and Planning Office can reveal the identifies of clients. In environments where collusion is possible between the Planning Office and the HMIS, additional safeguards are necessary to combat this threat (Section 12).

The Iowa Experiment posed a situation in which the community would likely rely on HMIS staff to perform the functions of the Planning Office, thereby introducing privacy risks due to possible collusion.

One problem is data linkage on demographics appearing in the shared client-level data. Section 11 reports that 36% of the Iowa clients had uniquely occurring combinations of {*year of birth*, *gender*, *5-digit ZIP*}, and the number jumps to 55% when including {*race*, *ethnicity*}.

1 Servicepoint is a product of Bowman Systems, servicing more than 30,000 clients in 45 states. They are a national leader in providing HMIS services. For more information, see http://www.bowmansystems.com/products.html.

Of course, just because a client has uniquely occurring demographics does not mean she is identifiable. The party seeking to re-identify data (termed "the linker" in this writing) must hold sufficient information to exploit this uniqueness. Section 2 reports that the likelihood in the USA of unique re-identifications of clients based on {*year of birth*, *gender*, *5-digit ZIP*} is only 0.04%. If the linker only has access to publicly available data (e.g., a voter list), then the likelihood a re-identification using these demographics is 0.04%. On other hand, if the linker is the HMIS in Iowa, which often contains non-domestic violence service records related to the same clients, then the likelihood of a re-identification using these demographics is about 36%.

One remedy to help thwart unwanted linking by the HMIS using demographic data elements is to only share the most general version of the data elements that still enable production of the AHAR. Section 11 reports that {first 3 digits of ZIP, gender, AHAR age ranges} was unique for 6% of the Iowa clients and was 11% when including {race, ethnicity}. This is a dramatic improvement, and even though it is not the only solution needed, sharing only the most general values lowers privacy risks overall.

A second problem is re-identification due to the linker exploiting the exact entry and exit dates appearing in the data (Section 11). A somewhat effective remedy is to replace exact dates of service with number of days of service or with time periods (e.g., overnight, 2-14 days, 15-30 days, 30 plus days). Section 11 provides more detail.

In order to prevent collusion when the Planning Office and the HMIS consist of the same personnel, it is necessary to use additional privacy safeguards, beyond those implemented in the PrivaMix Demonstration System or made possible by changes described to the Universal Data Elements. These safeguards involve post de-duplication anonymization. After PrivaMix de-duplication completes, additional processing would occur before it releases results to the Planning Office. Possibilities for post processing include: replacing client-level results with pivot tables that show aggregate count information for combinations of data elements; replacing client-level data with an overall final report (e.g., the AHAR itself); or, suppressing and generalizing outliers in the client-level results. Each of these approaches can provide additional and sufficient privacy protection, by replacing client-specific results with appropriately generalized ones. Section 12 describes these in detail.

In comparison to other approaches, the PrivaMix approach does not require domestic violence homeless shelters to share identifiable client data with a third party, a trusted third party, or an HMIS directly, as would a reporting service or centralized data storage, and provides better performance than encoding, hashing, encryption, scan cards, biometrics, and consent at constructing privacy-preserving UIDs. Section 9.9 and Figure 1 provides a comparison.

In conclusion, the PrivaMix Demonstration System achieved an accurate unduplicated accounting in the Iowa Experiment, and with the additional post processing anonymization described above, can do so while maintaining client privacy even in an environment in which the Planning Office and the HMIS are the same people.

More detailed information and recommendations appear below. These recommendations concern information collected from clients of domestic violence homeless shelters (termed "Clients" and "Shelters") and are not necessarily intended to be more generally applied to other homeless populations whose information may be captured in a HMIS. Figure 2, in Section 1.9, provides a quick summary of all recommendations.

1.1 General recommendations

Recommendation #1: Coordination of privacy protection schemes is necessary across planning offices that service a geographical region in which shelters within the region report to different planning offices but service some of the same clients. Lack of coordination can distort the unduplicated accounting. (For more information, see Section 3.3.)

Recommendation #2: A Shelter may assign a unique person identification number (PIN) to internally identify a client, but it should not share the client's PIN externally. PINs that include the Client's name, Social Security number, or other characteristic may be used alone or in combination with other data elements to re-identify a Client. Any characteristic not allowed as a data element or a UID, should not be used as an externally shared PIN. (For more information, see Section 3.5.)

Recommendation #3: If a Planning Office produces a De-identified Dataset from the HMIS data collected from Shelters, the De-identified Dataset should not include any original Personal Identification Numbers (PINs), Unique Identification numbers (UIDs), or Household Identification numbers. (For more information, see Section 3.6.)

Recommendation #4: A Shelter should release Client information to the Planning Office some time after the Client has left the shelter. (For more information, see Section 4.1.)

Recommendation #5: Shelters and planning offices should train personnel on the responsibilities and accepted practices for collecting, storing and sharing client information. (For more information, see Section 4.1.)

Recommendation #6: Unique Identification numbers (UIDs) values assigned to Clients of domestic violence shelters by Shelters should not be used (i.e., stored or referenced) by any non-HMIS program to which the Clients may participate in order to limit unwanted linking. For more information, see Section 4.2.)

Recommendation #7: Shelters and Planning Offices are already required to issue and post privacy notices to clients about the data collection, sharing, and linking practices of the shelters and planning offices in which the client's data will be part [1]. Beyond the role this requirement plays as a Fair Information Practice, this requirement is also important to help ensure the integrity of the information a client provides in forming the client's UID. (For more information, see Section 4.2.)

Recommendation #8: The fields *date of birth* and *ZIP code of last residence*, which are among the data elements Shelters share with Planning Offices, should contain information less specific than the month, day, and year of birth and all 5 digits of the ZIP (or postal) code. (For more information, see Section 4.5 and Section 7.1.)

Recommendation #9: A Planning Office may generate a "De-identified Dataset" from collected Shelter data to compute the unduplicated accounting. If so, the Planning Office should only use the Universal Data Elements in computing the De-Identified Dataset and remove (or obscure) elements from the De-identified Dataset that may appear in other data held by the Planning Office to limit secondary linking to other data held by the Planning Office. (For more information, see Section 4.6.)

Recommendation #10: Personnel in the Planning Office should sign a data use agreement with Shelters or provide notice to Shelters that either disallows the linking of the De-Identified Dataset to any other data or makes explicit the linking intended. (For more information, see Section 4.6.)

Recommendation #11: Given a "Proposed Solution" (i.e., a UID technology bundled with policies and practices for the construction, maintenance and use of a UID technology for clients of domestic violence homeless shelters), a person skilled in statistical, computational and/or legal principles, as appropriate, should certify in writing that the Proposed Solution has a minimal risk of re-identification when the solution is considered with other publicly and readily available information and techniques. Such writing should address vulnerabilities for inappropriate re-identifications by various categories of insiders. This is termed a "compliance statement" and should be made available for inspection. (For more information, see Section 5.5.)

Recommendation #12: Given a Proposed Solution, a person skilled in statistical and/or computational principles, as appropriate, should certify in writing that the Proposed Solution provides a reasonably accurate unduplicated accounting of client visit patterns to shelters within the regional setting it is to be deployed. Such writing should include possible false match and missed match rates. This statement is termed a "warranty" and should be made available for inspection. (For more information, see Section 5.5.)

1.2 Recommendations regarding UID technologies

The following recommendations result from assessments performed on the initial UID technologies explored by Shelters and Planning Offices. The list of initial technologies appear in Figure 1.

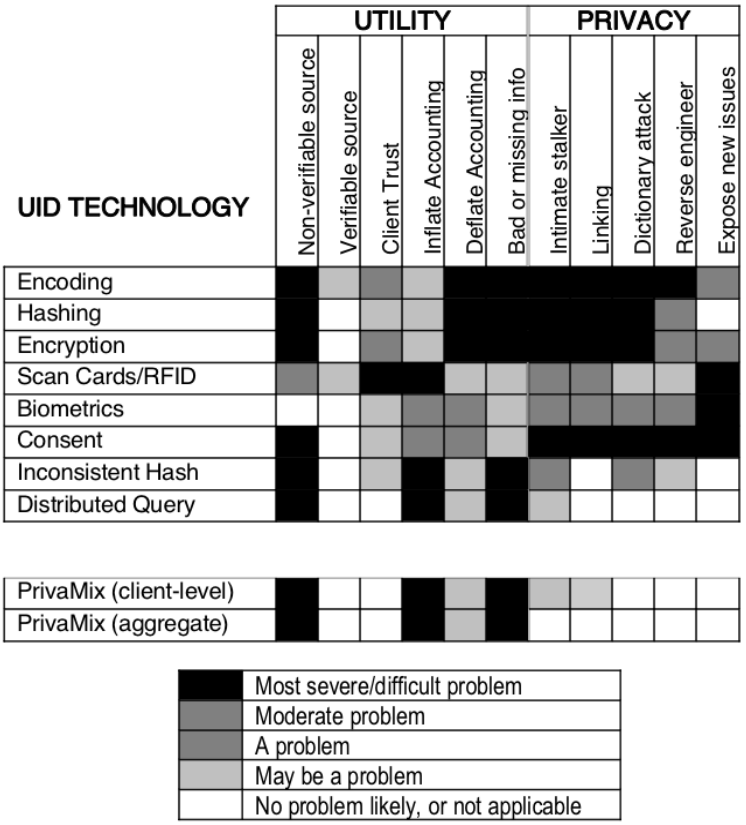

Figure 1. Technologies considered for UIDs. The top group are the initial technologies.

Recommendation #13: If the technology for constructing UIDs uses non-verifiable information from the client, then instruments that instill client trust in the overall system should be deployed; otherwise, the UID should use verifiable source input from clients. (For more information, see Section 6.9.)

Recommendation #14: If the technology for constructing UIDs involves encryption or hashing, then "strong" cryptographic methods should be used and the description of the method should be included in the warranty or compliance statement. (For more information, see Section 6.9.)

Recommendation #15: If the technology for constructing UIDs involves encryption or hashing, then accompanying practice should control access to and document an audit trail of specific uses of the encryption/hashing function. A description of these practices related to the capture and auditing of uses of the encryption/hashing function should be included in the warranty or compliance statement. (For more information, see Section 6.9.)

Recommendation #16: If the technology for constructing UIDs involves scan cards, then accompanying practices are needed to avoid issuing multiple cards to the same client and to prevent card sharing and swapping among clients. A description of practices related to avoiding these unwanted activities should be included in the warranty or compliance statement. (For more information, see Section 6.9.)

Recommendation #17: In cases where consistent UIDs are assigned to Clients over time, once Planning Offices link and de-duplicate Client visits, stored copies of the linked information should have all UIDs removed. (For more information, see Section 6.9.)

1.3 VAWA-based recommendations

In January 2006, Congress passed The Violence Against Women and Department of Justice Reauthorization Act of 2005, H.R. 3402 ("VAWA") which raised the privacy standard for UID technologies to guarantee clients cannot be re-identified. Recommendations related to the impact of VAWA on HMIS data elements and UID technologies appear below.

Recommendation #18: The fields *date of birth* and *ZIP code of last residence*, which are among the data elements in the Universal Data Elements, must contain information less specific than the month, day, and year of birth and all 5 digits of the ZIP (or postal) code in order to thwart linking. (For more information, see Section 4.5 and Section 7.1.) This is a strengthening of Recommendation #8.

Recommendation #19. The technology used to construct and de-duplicate UIDs must satisfy VAWA's requirements limiting re-identification. Consent and biometrics appear unable to satisfy the privacy standard established by VAWA. Encoding, hashing, and encryption may enable unwanted linking, and if so, pose grave concerns in attempts to use them to satisfy VAWA's privacy standard. Scan cards and RFID tags may be used, depending on the information appearing on (or within) the card. (For more information, see Section 7.2.)

1.4 PrivaMix recommendations

PrivaMix (Section 8) combines a form of inconsistent hashing (Section 6.7) with distributed query (Section 6.8) in three steps. These form the "PrivaMix Protocol."

The first step involves the assignment of UIDs. The same client gets different UIDs at different Shelters and can get the same UID at the same Shelter. This is done by using a strong one-way function having the commutative property. We term this a "PrivaMix function" (see Section 8.3 for requirements). Each Shelter computes a UID for a Client by applying the PrivaMix function to both a private value held by the Shelter and the source information provided by the Client (Section 8), thereby yielding different UIDs at different Shelters for the same Client.

In the second step, Shelters ship Client visit information to the Planning Office. Each record contains requested visit information and the Client's UID. At the end of this step, the Planning Office has visit details for all Clients at all Shelters, but does not know which UIDs relate to the same Clients across Shelters.

In the third step, Shelter and Planning Office machines communicate over a network to de-duplicate UIDs. We term the network of machines, a "PrivaMix Network." Each Shelter applies the PrivaMix function to its private value and the UIDs from all the other Shelters once in a process we term "mixing." After all Shelters finish mixing, complete mixes for UIDs will only

be the same if the original Client source information was the same. This identifies which UIDs refer to the same Client.

There are variations to the generic PrivaMix Protocol to address particular issues.

PrivaMix Variation 1: Shelters mix among themselves, without the Planning Office (Section 8.2.2).

PrivaMix Variation 2: Shelters check that UIDs are legitimate (Section 8.2.3).

PrivaMix Variation 3: Matching UIDs to Universal Data Elements (Section 8.2.4).

PrivaMix Variation 4: Providing aggregate count distributions, not Client-level data (Section 8.2.5).

PrivaMix Variation 5: Anonymizing client-level data (Section 8.2.6).

PrivaMix Variation 6: Using web browsers for mixing (Section 8.2.7)

Recommendations below relate to using PrivaMix as a UID technology.

Recommendation #20: When using PrivaMix as a UID technology, care should be taken to avoid multiple Shelters from having the same private value. The Shelter's private value customizes the PrivaMix function to the Shelter. If multiple Shelters inadvertently have the same private value, then those Shelters assign exactly the same UIDs to the same clients. In most uses of PrivaMix, the UIDs will only be used for one-time mixing. In these cases, it is okay if Shelters inadvertently select the same private value though the likelihood of such should be rare. (For more information, see Section 8.1.)

Recommendation #21: When using PrivaMix as a UID technology, if the visit data is transmitted to the Planning Office over the PrivaMix network of Shelter and Planning Office machines, then appropriate computer security standards for the storage of Client information should be enforced because these machines contain Client source and visit information. (For more information, see Section 8.1.)

Recommendation #22: If desirable, use a variation of the PrivaMix Protocol to have a party other than the Planning Office orchestrate mixing. One variation (Section 8.2.2) describes how Shelters perform mixes among themselves and then forward de-duplicated results to the Planning Office. Another variation (Section 8.2.7) describes how a third-party might orchestrate de-duplication and then forward results to the Planning Office.

Recommendation #23: Thwarting data linkage threats requires further privacy consideration, realized as variations of PrivaMix and/or dictates on data elements. Rather than PrivaMix providing Client-level data to the Planning Office, PrivaMix can alternatively provide aggregate de-duplicated count distributions (Section 8.2.5). A way to help thwart data linkage threats within PrivaMix while still providing Client-level data is to anonymize the data after de-

duplication (Section 8.2.6). An alternative that lies outside of PrivaMix is to chose non-identificable Client-level data elements (Section 11).

Recommendation #24: An economical implementation of the PrivaMix Protocol involves using traditional web browsers already provided with computers (Section 8.2.7). Doing so has the advantage that no dedicated machine is needed, that no additional software has to be installed, and that no intense user training is needed.

Recommendation #25: A PrivaMix function (**F**) must satisfy the following six requirements (Section 8.3):

(1) Inconsistent assignment: different shelters should generate different initial mix values for the same clients.
(2) One-way function: **F** must be a one-way function.
(3) Commutative: **F** must be a commutative cipher.
(4) Privacy: the secret client information cannot be learned given the sharing of complete and sub-mixes.
(5) Collision-free: mixes from **F** must be collision-free.
(6) Correctness: all complete mixes for the same client must be the same. Complete mixes for different clients should not be the same.

Here are seven statements claimed about PrivaMix. These form the basis of the recommendations that follow them.

Usability claim. Communication time is linear in the number of Shelters. (Section 8.4.1.)

Correctness claim. If the complete mixes are the same, the Clients representing the original UIDs presented the same source information.. (Section 8.4.2.)

Privacy claim. A dictionary attack by the Planning Office will not yield reliable re-identifications. (Section 8.4.3.)

Privacy claim. Compromising a Shelter will not help the intimate stalker learn where a targeted Client is (or has been). Similarly, compromising the Planning Office will not help the intimate stalker learn where a targeted Client is (or has been). (Section 8.4.4.)

Privacy claim. Even if the Planning Office pads the UIDs with known values, the Planning Office does not learn Client source information. (Section 8.4.5.)

Limitation. If the Planning Office and at least one Shelter collude, the Planning Office can learn Client source information about the Shelter's Clients and the Shelter can learn other Shelters its Clients visited. (Section 8.4.6.)

Limitation. If during the de-duplication protocol, the intimate stalker compromises both the Planning Office and a Shelter the targeted Client visited, the intimate stalker can learn the locations of all Shelters the Client visited. In addition, the Planning Office can learn the source information for that Client. (Section 8.4.7.)

Recommendation #26. Each Shelter must select a sufficiently private value so that efforts by the Planning Office to exhaustively compute all combinations of Shelter private values and Client source information (a dictionary attack) are not feasible. Most likely a Shelter's computer will be required to select a private value 512 bits or larger as appropriate and most likely randomly selected at the start of each reporting period. (For more information, see Section 8.4.)

Recommendation #27: To help thwart the possibility of the Planning Office or other Shelters learning a Shelter's private value, a Shelter may not even explicitly know its own private value for a reporting period –i.e., the computer program may generate it internally and not explicitly reveal it. (For more information, see Section 8.4.)

Recommendation #28: To help thwart the possibility of the Planning Office or other Shelters learning a Shelter's private value, a Shelter may make its private value available to its copy of the PrivaMix function only while mixing over the PrivaMix Network. Other parties should not be able to invoke a Shelter's PrivaMix function with the Shelter's private value. (For more information, see Section 8.4.)

Recommendation #29: In order to prevent the Planning Office from padding UIDs with known values, the original PrivaMix approach should be modified to validate the number of UIDs and/or to mix UIDs without Planning Office involvement. (See Variation 1 and Variation 2 in Section 8.2 for details and Section 8.4 for motivation.)

Recommendation #30: Care must be taken to combat possible collusion between the HMIS and the Planning Office because in many geographical regions, the staff of the HMIS is the same staff as the Planning Office (or CoC) and because there is a desire to de-duplicate visits across the domestic violence homeless shelters and the HMIS (not the domestic violence homeless shelters alone). As a participant in PrivaMix, a HMIS poses a significant threat to Client re-identifications because a HMIS will usually contain most (if not all) Clients who visit any domestic violence homeless shelter. Remedies include having PrivaMix provide only aggregate information or provably anonymizing released data elements. (See Section 12 for details and Section 8.4 for motivation.)

Recommendation #31: Client records Shelters provide to the Planning Office should only include Clients who are no longer residing at the Shelter. This is a helpful recommendation, but not wholly satisfactory because Clients may re-visit previously visited Shelters. (For more information, see Section 8.4.)

Recommendation #32: The Planning Office should destroy all copies of the original UIDs once the de-duplication is complete. Doing so, limits the opportunity for compromise. (For more information, see Section 8.4.)

Recommendation #33: A specific implementation of a system that uses the PrivaMix approach requires revisiting claims and limits specific to implementation details. Differences in implementations may include communication flow (e.g. Planning Office in the middle or Shelter-to-Shelter), information content (e.g., a stream of values, or a list of values with their originating Shelter), and selection of the privately held Shelter value (.e.g., random selection, or pre-selection). (For more information, see Section 8.4.)

In comparing PrivaMix with the UID technologies discussed earlier, PrivaMix performs comparable to inconsistent hashing (Section 6.7) and distributed query (Section 6.8) making it generally better than encoding (Section 6.1), hashing (Section 6.2), encryption (Section 6.3), scan cards and RFIDs (Section 6,4), biometrics (Section 6.5), and consent (Section 6.6) at protecting privacy. Yet, the utility of its de-duplicated results is better than encoding, hashing, encryption, scan cards and RFID, but not better than biometrics or consent. (For more information, see Section 8.5.)

1.5 The PrivaMix Demonstration System

In 2007, Privacert implemented a version of PrivaMix for a real-world experiment; we term this software the "PrivaMix Demonstration System." Here is a quick summary of its highlights.

- uses regular computers operating over the Internet
- each participant (Shelter and Planning Office) has its own machine
- data is shared using standard comma-delimited text files
- the Planning Office machine coordinates mixing
- final de-duplicated results don't include UIDs or complete mixes, just sequential numbers

Because there are numerous variations and many ways to implement the PrivaMix Protocol, Section 9 describes the details of the PrivaMix Demonstration System specifically. Section 10 explains its use in the real-world experiment. Below is a brief description of the PrivaMix Demonstration System.

In the PrivaMix Demonstration System, each participating machine runs special software devoted to this task. Shelter machines run one edition of the software program ("the Shelter Edition"). The Planning Office machine runs a different edition ("the CoC Edition"). These editions differ because the responsibilities of Shelters and the Planning Office in the PrivaMix protocol are different. (For more information, see Section 9.)

Operation of the PrivaMix Demonstration System is extremely simple. If Shelters and the Planning Office use default settings, then operation is as simple as loading the Client information and clicking one button. (For more information on user options and screen shots, see Appendix A.)

The PrivaMix Demonstration System has minimal machine requirements, which means almost any computer system sold today is sufficient for use. However, the machine must have access to the Internet. (For more information, see Section 9.1.)

The Shelter provides an initial comma-delimited text file for processing, which has the fields that comprise the Client's source information appearing as the leftmost fields. The remaining fields on the line are fields associated with the Client's visit to the Shelter, presumably the Universal Data Elements associated with that Client. After the Shelter machine computes UIDs for each Client from Client source information, it produces a comma-delimited file replacing the leftmost fields with Client UIDs. Shelter machines then transfer the resulting comma-delimited text file to the Planning Office as encrypted content over an Internet connection. (For more information, see Section 9.5.)

While the PrivaMix Demonstration System does not dictate which Client fields to use as source information, precautions are needed. Below are two important precautions. (For more information, see [32] and Section 9.4.)

1. Care must be taken that sufficient variability exists in the fields so that resulting UIDs have a sufficiently wide range of possible values.

2. Care must also be taken to make sure that different Clients are not likely to have to the same set of values appearing in the source information.

In the PrivaMix Demonstration System, the Planning Office orchestrates mixing as described in the generic PrivaMix Protocol (Section 8.2). The Planning Office sends values to each Shelter, one Shelter at a time, to mix, such that each Shelter mixes each UID once.

After mixing completes, the PrivaMix Demonstration System performs de-duplication on the Planning Office machine matching complete mixes across Shelter data. All values are held in the computer's memory. No information appears on the hard drive. (For more information, see Section 9.7.)

Before making final de-duplicated results available to the Planning Office, the PrivaMix Demonstration System removes all UIDs, replacing them with numbers from 1 to the total number of distinct Clients. The Planning Office does not receive a copy of the UIDs or complete mixes, only the results of de-duplication. (For more information, see Section 9.8.)

1.6 The Iowa experiment

In a real-time experiment with three shelters, an HMIS and a Planning Office, a "PrivaMix Demonstration System" computed an accurate unduplicated accounting using real-world data from homeless programs in Des Moines, Iowa ("the Iowa Experiment"). Here is a summary of experimental results. For details, see Section 10.

The experiment used laptops with wireless broadband network, with the software loaded and pre-configured for operation. Standardizing the machines allowed the experiments to focus efficiently and narrowly on performance.

Subjects were clients whose data appeared at participating shelters and the HMIS in a previous six-month time period. The actual subjects are not clients of domestic violence ("DV") homeless shelters, but are clients of homeless family shelters (not domestic violence specific). Using non-DV shelters allowed us to compare computed de-identified results with results derived manually using fully identified data. Of course, the generalizability of these experiments assume there is no difference between DV and non-DV data collection.

A key component in de-duplicating UIDs is the Client source information used to construct the UIDs. Fields having omissions or errors can render UIDs useless. While the PrivaMix Demonstration System works with any Client source information, Privacert proposed to use the first three letters of the first name and the date of birth. Experiments compared Privacert's proposed method with using Social Security numbers, and two methods currently in use by

Servicepoint. Privacert's method encountered fewer fields having omissions or errors than the other methods, and used fields in which clients provided more consistent values than the fields used by the other methods. In performing an unduplicated accounting, the Privacert method had the lowest number of errors.

After constructing UIDs, shelters, the HMIS, and Planning Office conducted a real-time duplication using the laptops located at their facilities. The PrivaMix Demonstration System performed exactly as if plain text was used even though sensitive Client source information was provably never shared with the Planning Office or the other Shelters. No errors were introduced.

1.7 Changes to the Universal Data Elements

The generic PrivaMix approach solves privacy and utility problems related to the assignment and de-duplication of UIDs. However, privacy threats may remain from data linkage capabilities afforded by the Universal Data Elements. Below are recommendations related to demographics appearing in the Universal Data Elements.

Recommendation #34: The AHAR does not require the demographic specificity currently found in the Universal Data Elements. More general values can be shared without any loss to reporting ability. Therefore, the Universal Data Elements should be revised to reduce the likelihood of recognition by the intimate stalker and/or data linkage threats by using the most general values possible. (For more information, see Section 11.)

Recommendation #35: The *date of birth* field should minimally be an *age range*. In fact, a Client may have more than one kind of age range specification. For example, there may be a data element related to 5-year age ranges, and another related to AHAR ranges (under 1, 1 through 5, 6 through 12, 13 through 17, 18 through 30, 31 through 50, 51 through 61, and 62 and over), enabling more reporting uses of the resulting data. (For more information, see Section 11.)

Recommendation #36: The *ZIP of last residence* field should be changed to either report the *first 3 digits of ZIP*, or even better, be changed to be a boolean flag denoting whether the Client's last residence was *within the geography covered* by the Planning Office or not. If the *first 3 digits of ZIP* are used, then only those values local to the Planning Office need be recorded. Clients from outside the local area would just have a special value, like 999, in order to prevent them appearing as unique outliers. (For more information, see Section 11.)

Recommendation #37: *PIN* should be removed. The Shelter should not provide its internal unique number. Instead, the Shelter should maintain an exact copy of the data provided so that records can be referred to in discussion with the Planning Office by the place (or row) in which the record appears. (For more information, see Section 11.)

Recommendation #38: Consider removing *Race* and *Ethnicity*. Experimental results showed that the addition of these fields increase risks to re-identification. (For more information, see Section 11.)

Recommendation #39: Shelters should consider renumbering *Household identification numbers* from 1 to the last household, prior to forwarding the information to the Planning Office. This makes sure the household identification number itself cannot be the basis for linking. (For more information, see Section 11.)

Recommendation #40: Replace the exact service dates (*Program Entry Date* and *Program Exit Date*) with number of days of service or with time periods (e.g., overnight, 2-14 days, 15-30 days, 30 plus days). (For more information, see Section 11.)

Recommendation #41: More sensitive data elements (such as *first name*, *Social Security number*, or full *date of birth*) may still be collected by Shelters in order to produce a useful UID. However, those values should continue to not be forwarded to the Planning Office as part of the Universal Data Elements. (For more information, see Section 11.)

1.8 Privacy assurance recommendations

In order to prevent collusion when the Planning Office and the HMIS consist of the same personnel, it is necessary to use additional privacy safeguards, beyond those for protecting UIDs (e.g. PrivaMix) and beyond merely changing the Universal Data Elements. Remedies involve expanding the post-processing done by PrivaMix so that the final dataset made available to the Planning Office contains either aggregate (not Client-level data) or provably anonymized Client-level data.

While PrivaMix guarantees privacy protection for UID creation and use in de-duplicating, linking vulnerabilities currently remain in the de-duplicated Universal Data Elements (Section 11). Problems stem from the selection of which data elements to associate with UIDs, and not from the UIDs themselves. Changes to the Universal Data Elements can help (Section 11), but such changes seem unable to be wholly satisfactory without effecting the usefulness of the de-duplicated data to the AHAR.

A PrivaMix System can anonymize de-duplicated results prior to forwarding data to the Planning Office. The anonymizaed data will not be vulnerable to linking, even if the Planning Office and HMIS collude.

At present, the PrivaMix Demonstration System, as used in the Iowa Experiment, de-duplicates Client information and then passes values associated with each UID to the Planning Office "as is." Instead of merely forwarding those values, a PrivaMix System could anonymize those data elements and then forward the anonymized results to the Planning Office.

There are numerous way to perform the anonymization. These include: replacing client-level results with pivot tables that show aggregate count information for combinations of data elements; replacing client-level data with an overall final report (e.g., the AHAR itself); or, provably anonymizing client-level data by automatically suppressing and generalizing values as needed. Each of these approaches can provide sufficient privacy protection, by replacing client-specific results with appropriately generalized ones. The result is privacy protection, even

against data linking, and accurate de-duplicated results for the AHAR. (For more information, see Section 12.)

Recommendation #42: In order to prevent collusion when the Planning Office and the HMIS consist of the same personnel, it is necessary to use additional privacy safeguards, beyond those for protecting UIDs (e.g. PrivaMix) and beyond merely changing the Universal Data Elements. It is necessary to make sure the HMIS cannot link the Universal Data Elements to other service information contained in the HMIS. (For more information, see Section 11.)

Recommendation #43: Add post de-duplication anonymization to a PrivaMix System to make sure data provided to the Planning Office is not vulnerable to linking, even if the Planning Office and HMIS collude. The Planning Office receives provably anonymized de-duplicated results. (For more information, see Section 12.)

Recommendation #44: Consider having the final results be aggregate data only. Instead of Client-level data, a PrivaMix System can alternatively provide aggregate de-duplicated count distributions denoting how many Clients matched particular characteristics. An example of a count distribution are counts by age ranges. Distributions can involve more than one field to get more specific data. (For more information, see Section 8.2 and Section 12.)

Recommendation #45: Consider having the final results be the AHAR report itself. Instead of Client-level data, a PrivaMix System can alternatively provide the AHAR to the Planning Office. (For more information, see Section 8.2 and Section 12.)

Recommendation #46: Consider having the final results be anonymized Client-level data. Anonymized Client-level data generalizes or suppresses values, as needed, to protect privacy. Formal protection models identify which values to generalize or suppress from the resulting dataset so that each record ambiguously relates to a minimum number of people [30][31]. For example, if a 80 year old woman is an outlier in the data because of her age, either her age would be removed from the data or generalized to a category having more people, such as "50 plus" as appropriate value given the other ages appearing in the data. (For more information, see Section 8.2.6 and Section 12.)

In conclusion, PrivaMix provides an effective and accurate privacy-preserving means for constructing and de-duplicating UIDs. However, additional care with the Universal Data Elements must be taken to properly protect against unwanted data linkage with the HMIS. The problem is not with the UIDs but with the selection of data elements associated with the UIDs. A solution is to enhance a PrivaMix System to anonymize de-duplicated Client-level data and then forward the anonymized results to the Planning Office.

1.9 Summary of recommendations

Figure 2 below contains a quick summary of recommendations made. Some recommendations repeat because of the context in which it appears in the text.

#	Description	Section
1	Coordinate de-duplication across neighboring CoC's.	3.3
2	Not share Shelter PIN beyond Shelter.	3.5
3	De-duplicated results should not include PINs, UIDs, or Household IDs.	3.6
4	Shelters only include Clients who have left the Shelter.	4.1
5	Train personnel on accepted practices for handling Client data.	4.1
6	UIDs should be inconsistently assigned across Shelters.	4.2
7	Shelters should privacy notices for Client inspection.	4.2
8	Fields date of birth and ZIP should be less specific.	4.5, 7.1
9	Planning Office should delete any fields in the Universal Data Elements not needed.	4.6
10	Planning Office should sign Data Use Agreement with Shelters regarding linking.	4.6
11	Skilled person should certify System's risk of re-identification.	5.5
12	Skilled person should certify utility of de-duplicated results.	5.5
13	System using non-verifiable source information should instill trust.	6.9
14	System using encryption or hashing should use strong cryptographic methods.	6.9
15	System using encryption or hashing should control access to the function.	6.9
16	System using scan cards/RFID should avoid issuing multiple cards to the same Client.	6.9
17	UIDs should be removed from de-duplicated results.	6.9
18	Fields date of birth and ZIP must be less specific.	4.5, 7.1
19	System must satisfy VAWA's requirements limiting re-idenification.	7.2
20	A PrivaMix System must avoid Shelters producing the same UID for Clients.	8.1
21	Computers transmitting UDE over a network must adhere to accepted security standards.	8.1
22	If desirable, have a party other than the Planning Office orchestrate mixing.	8.2
23	A PrivaMix System should anonymize or aggregate results, rather than provide Client-level data.	8.2, 11
24	An economical PrivaMix System can result from using existing web browsers.	8.2
25	A PrivaMix Function must satisfy six noted requirements.	8.3
26	In a PrivaMix System. A Shelter value must be sufficiently large.	8.4
27	In a PrivaMix System, a Shelter should not even know its own private value.	8.4
28	In a PrivaMix System, unauthorized parties should be unable to use the Shelter's PrivaMix function.	8.4
29	In a PrivaMix System, Shelters should validate the number of UIDs requested to mix.	8.2, 8.4
30	In order to provide collusion with an HMIS, provide only aggregate or anonymized results.	8.4, 12
31	Shelters only include Clients who have left the Shelter.	8.4
32	UIDs should be removed from de-duplicated results.	8.4
33	Claims must be assessed for any particular PrivaMix implementation.	8.4
34	Make Universal Data Elements as general as remains useful to the AHAR.	11
35	Make date of birth field more general, such as the AHAR age classifications.	11
36	Make ZIP of last residence field more general, such as a boolean flag denoting whether in covered	11

#	Description	Section
	area.	
37	Remove PIN field from the Universal Data Elements.	11
38	Consider removing race and ethnicity fields from the Universal Data Elements.	11
39	Consider having Shelters renumber Household IDs to thwart any possible linking using the field.	11
40	Replace exact service dates with number of days or time periods.	11
41	Sensitive data elements may be used for UIDs, but not forwarded to the Planning Office.	11
42	Use privacy protections beyond UIDs and modified Universal Data Elements to thwart linking to HMIS.	11
43	Consider PrivaMix performing post de-duplication anonymization to thwart linking to HMIS.	12
44	Consider PrivaMix providing aggregate values, not Client-level data, to the Planning Office.	8.2, 12
45	Consider PrivaMix providing the AHAR itself, not Client-level data to the Planning Office.	8.2, 12
46	Consider PrivaMix providing anonymized Client-level data to the Planning Office.	8.2, 12

Figure 2. Summary of recommendations.