4. Privacy Threats

There are two primary motivations for infringing on the privacy of Clients of domestic violence shelters: the intimate abuser seeks to learn the physical location of the Client; and, the Planning Office seeks to link Client information to other available data to learn more about Clients overall. The next sections discuss these threat models in detail.

4.1 Intimate stalker threat

Domestic violence shelters have historically had to protect Clients from intimate and aggressive abusers and concerns are well founded. Over 31% of all women[7] murdered in the United States are murdered by husbands, boyfriends , or exes – the majority killed after attempting to leave an abusive relationship [16][17]. The National Institute of Justice estimated that 73% of domestic violent assaults go unreported largely because of women's lack of faith in the system [17].

Personal stories are quite chilling. As an example, consider a case from Los Angeles, California [18]. In 2001, a woman's husband was unemployed and had been drinking heavily. When she refused to have sex with him, he attacked her, prevented her from calling for help, and held her captive in her home. Various other incidents recurred. Eventually she was able to get a spot in a family shelter for herself and her two children. After leaving the shelter, the husband quickly tracked her down and strangled her to death with a belt.

The "intimate stalker" (an name given in this writing to an intimate abuser who stalks a Client) challenges computer systems that record and share Client visit information in several ways. First, the intimate stalker typically has knowledge of various personal facts about the Client that may be recorded in data held by the Shelter in which the victim resides. For example, an intimate stalker is likely to know the victim's name, date of birth and Social Security number, which may not be readily known by the general population. Second, the intimate stalker tends to be highly motivated to locate a targeted Client. For example, repeated violations of court orders and police reports describing escalating incidents of death threats, stalking and harassment are common. Finally, an intimate stalker may use insider access (either his own or by compromising an insider who has access to the data) to gain location information on a targeted Client. For example, an intimate stalker may persuade a family member or a friend to assist in revealing a Client's Shelter location by expressing a desire to reconcile for the sake of children or because situations (such as obtaining a new job) have changed.

No one solution addresses all these concerns, however some recommendations can be made immediately and others will be made in subsequent sections.

7 While the wording used has a bias that women are victims and men are abusers, it is important to note that men are also victims and that abusers can be male or female.

One recommendation, as stated below, is to thwart the intimate stalker's ability to locate the Client by making sure visit information shared with the Planning Office is no longer current. This protection is not a first line of defense against an intimate stalker and should not be the only protective action taken. It merely offers supplementary protection. Stronger protections, which will be examined later in this writing, guarantee that the location of any Shelter in which the Client has historically visited cannot be learned by the intimate stalker. Stronger protection is important because some Clients tend to re-visit the same Shelters and an intimate stalker's knowledge of a historic visit can pose future problems.

Recommendation #4: A Shelter should release Client information to the Planning Office some time after the Client has left the shelter.

Another recommendation, as stated below, is aimed at helping thwart the stalker's ability to recruit or compromise those with insider access to Client information. This protection only provides supplemental protection. Stronger protections, which are examined later in this writing, guarantee that the Client's information cannot be found in information shared or stored external to the Shelter.

Recommendation #5: Shelters and planning offices should train personnel on the responsibilities and accepted practices for collecting, storing and sharing client information.

4.2 Data linkage threat

Beyond the intimate stalker threat in which information about a single Client is sought, the data linkage threat involves learning information about most, if not all, Clients by matching the information to other available data in order to use HMIS data inappropriately. This kind of activity is most likely to occur at Planning Offices where linking can be used to learn information about a larger number of Clients than those at just one Shelter. Protecting privacy in this setting cannot involve thwarting all linking, because the HMIS de-duplication task the Planning Office performs on the data requires linking records that belong to the same Client across Shelters. Instead of thwarting all linking, privacy protection in the HMIS setting involves thwarting linking attempts that may re-identify Clients.

Figure 10 provides an example in which the Dataset is linked to publicly available voter information on {*ZIP*, *date of birth*, *sex*} to re-identify the records in the Dataset by *name*. The more uniquely occurring {*ZIP*, *date of birth*, *gender*}, the more fruitful the re-identifications.

Most UIDs are designed to be uniquely assigned to Clients, so as a result, UIDs can also be used as the basis for linking datasets. That is not surprising given that HUD introduced UIDs into HMIS in order to link Client visits. However, if the same UIDs are also used with non-HMIS data, then they become the basis for linking HMIS data beyond the HMIS context. The following recommendation is aimed at thwarting secondary uses of HMIS data using UIDs.

Recommendation #6: UID values assigned to Clients of domestic violence shelters should not be used (i.e., stored or referenced) by any non-HMIS program to which the Clients may participate to limit unwanted linking.

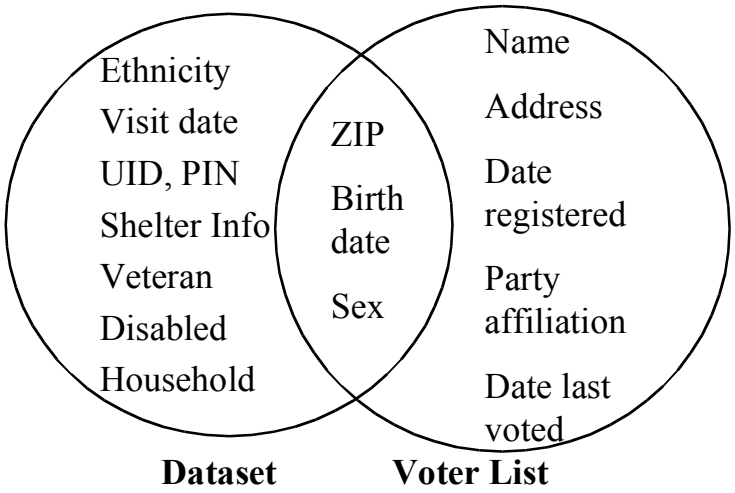

Figure 10. Example of linking Dataset to a publicly available population register, such as voter list, to re-identify the names of Clients appearing in Dataset.

In some cases, Planning Offices may decide to use HMIS data outside the HMIS context and in so doing, may purposefully link HMIS data to other non-HMIS data, even though this is unnecessary to achieve HMIS objectives. UID technologies can be constructed to thwart this behavior, as discussed later in this writing, but if this activity is desired, then Clients and Shelters should be made aware of this practice and any increased risk that may result. This is the motivation behind the following recommendation.

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.

4.3 Re-identification

A "re-identification" results when a record in Dataset can reasonably be related to the Client who is the subject of the record in such a way that direct and rather specific communication with the Client is possible. Figure 11 provides a depiction of a re-identification in which external information is linked on month and year of birth (9/1960), gender (F), and ZIP code (37213) to identify the visit information as belonging to Ann. The re-identification is sufficient to send a letter to Ann's residence.

For another example, consider Figure 10 in which Dataset is linked to a voter list to re-identify Client visits by name, even though Client names had been omitted from the visits in an attempt to protect privacy (recall Section 3.4.3).

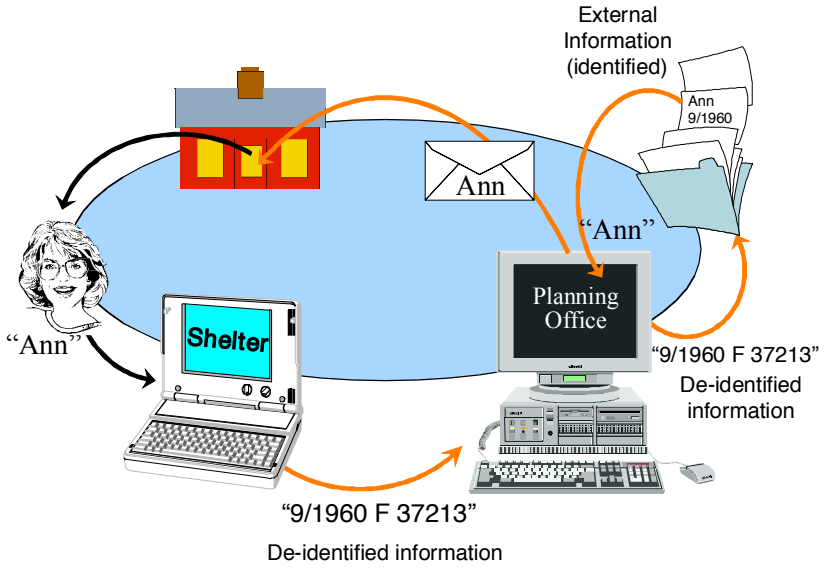

Figure 11. Depiction of re-identification. Ann leaves her home and gives her explicitly identified information to the Shelter. De-identified information about Ann is provided to the Planning Office, but in this depiction, the information can be used with external information (or personal knowledge) to re-identify the information as belonging to Ann. A re-identification occurs if there is sufficient information to directly communicate with Ann (not limited to mail), shown in the diagram as mailing an envelope to her original residence (or alternatively, sending the letter to Ann at the Shelter in which she resides).

4.4 Identifiability

One way to report the risk of re-identification is to determine the number of people to whom a record could refer. This is termed "identifiability." Figure 12 shows two examples in which information is released and compared against a known population. On the left, Figure 12 (a), each of the released profiles are ambiguous in terms of head shape and shading. Neither can be uniquely identified. The top released profile matches Hal and Len indistinguishably and the bottom profile ambiguously matches Jim and Mel. The release shown on the upper right of Figure 12 (b) is different. There is only one person in the known population (Hal) having the same color and head shape. In this case, the record referring to Hal is uniquely re-identified even though many of Hal's details had been removed.

While unique re-identifications obviously pose a privacy problem, so do situations in which a record maps ambiguously to a few known people. In Figure 12(a), both released profiles map to two individuals, but these people are both explicitly known, so they can both be contacted with little effort. Of course, the larger the number of people to whom a record refers, even if all of the people are known, the greater the effort usually needed to contact so many or make use of the information.

Counting the number of possible re-identifications for a record is a useful measure of privacy risk, but what is needed is a way to estimate the number of people to whom a record might refer.

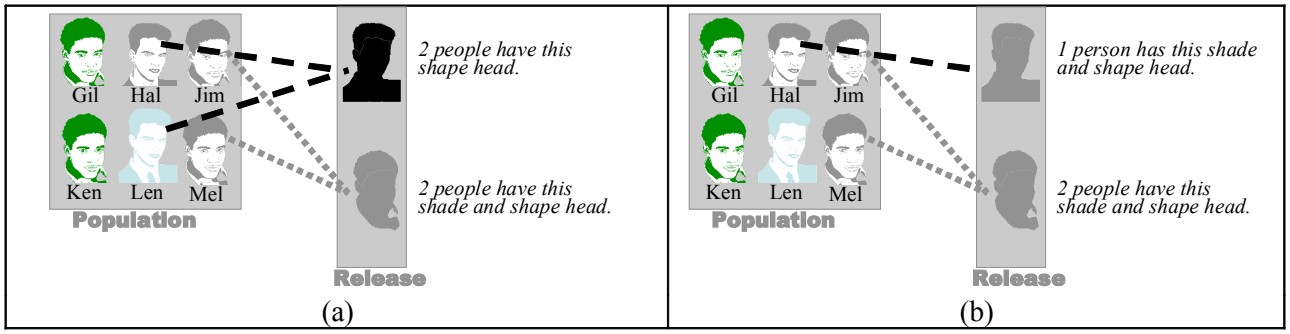

Figure 12. The identifiability of the profiles released in (a) are each ambiguously re-identified to two named persons. The top profile released in (b) is uniquely re-identified to Hal.

4.5 Identifiability of a dataset

The Risk Assessment Server is a commercially available system that reports re-identification risks by estimating the number of named persons to which each record could relate given its model of the U.S. population and its knowledge of publicly available datasets [20]. The output of the Risk Assessment Server is a plot of identifiability estimates, in graduated size groupings, that report the number of people to which a released record is apt to refer.

Figure 13 shows the results from the Risk Assessment Server based on {*date of birth*, *gender*, *5-digit ZIP*} from Dataset. The lower left plot shows that 87% of the population are uniquely identified by these characteristics. As age information is generalized and as geographical reference to the Client's prior residence is made less specific, uniqueness deteriorates and privacy protection increases. For example, {*year of birth*, *gender*, *5-digit ZIP*} drops the unique identifiability to 0.04% (see the lower right plot in Figure 13).

Dataset currently requires Shelters to provide the full month, day and year of birth and all 5 digits of the Client's last residential ZIP code, yet the AHAR uses only gross age values and geography relative to Shelter's service area (refer to Section 3.6). The following recommendation is aimed at increasing privacy protection by changing the level of specificity in these fields.

<u>*Recommendation #8:*</u> *The fields* date of birth *and* ZIP code of last residence*, which are among the data elements HUD recommends HMIS collect in the Universal Data Elements, should contain information less specific than the month, day, and year of birth and all 5 digits of the ZIP (or postal) code.*

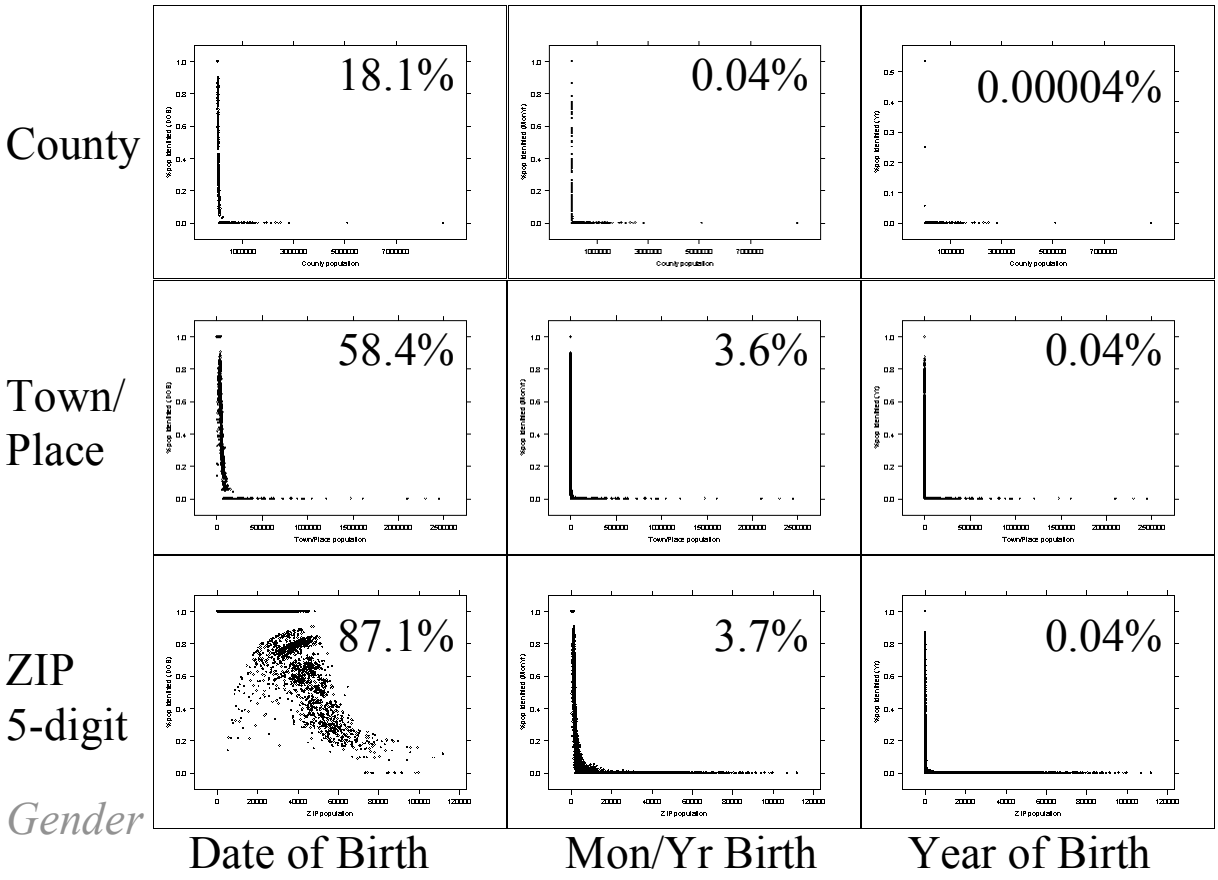

Figure 13. {*date of birth*, *gender*, *5-digit ZIP*} uniquely identifies 87.1% of USA population, but as ZIP is made less specific, the identifiability drops to 18.1% (bottom to top). Similarly, as the age of the client is made less specific, the identifiability drops to 0.04% (left to right). All values include gender. The horizontal axis of each sub-plot is the number of people who reside in the geographical area and the vertical axis is the percentage of the population uniquely identified by the noted combination of demographics noted. As the demographics are aggregated, the points move towards 0% identifiable.

4.6 Privacy concerns in Program-Specific Data Elements

Planning Offices that receive Program-Specific Data Elements (Figure 6) have some additional privacy concerns to consider to best protect Client data.[8] Program-Specific Data Elements may be linked to other available programmatic information to re-identify Clients. This vulnerability differs among municipalities and states as different kinds of secondary data from related programs are available.

A Planning Office is assumed to have multiple versions of data available, each having different re-identification risks and therefore different access policies. Figure 14 provides an overview. In terms of re-identification risk, the most sensitive data is that which first arrives at the Planning Office from the Shelter. These data may be separated into the Dataset used for the unduplicated accounting (the Universal Data) and the Program-Specific Data. No UIDs should appear in the Program-Specific Data. The De-identified Dataset is of least risk. A Planning Office may make internal access policies commensurate with these levels of risk. This advice regarding the maintenance of various versions of data is for consideration by Planning Offices and is not required.

Different versions of the data have different purposes. The originally received data could be maintained intact for quality control of Client information with Shelters (using PINs). The De-identified Dataset (modified to have less specific values of ZIP and date of birth) offers the least risk of re-identification and can be used to compute the unduplicated count information. In cases where the Shelter does not provide Program-Specific Data, the Dataset and the Originally Received Data are the same.

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.

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.

8 The requirements of the Program-Specific Data elements reside outside the scope of this work. However, some relative re-identification risk is noted.

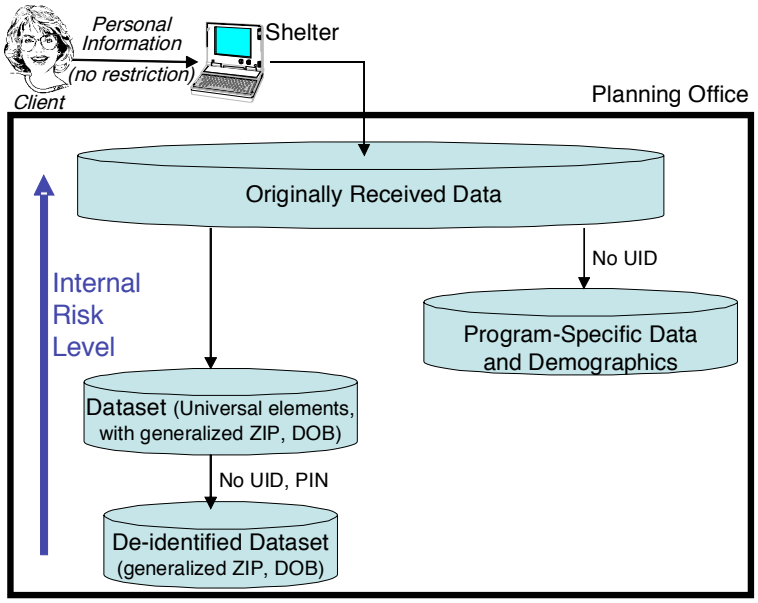

Figure 14. Versions of data maintained by a Planning Office with relative internal risk of re-identification. The originally received data has the most internal risk and the De-identified Dataset has the least.