12. Privacy Assurance Using PrivaMix

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 for a PrivaMix System to perform 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.

A way to thwart HMIS linking of Universal Data Elements without expanding PrivaMix is to have all clients, whether they be domestic violence clients or not, use the same privacy protections of the domestic violence clients. Then, the HMIS itself would lack explicit identifiers of clients, making linking less useful. The viability of this option in terms of the overall utility of the HMIS is beyond the scope of this writing.

Below are recommendations based on the discussion above.

<u>Recommendation #43:</u> *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.*

Recommendation #44: 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.

Recommendation #45: 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.

Recommendation #46: 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.

Recommendation #47: 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.

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.

Acknowledgements

The author thanks Michael Roanhouse, Mary Joel Holin, Julie Hovden, Brian Sokol, Michelle Hayes, Julie Eberbach, David Eberbach, Eileen Mitchell, Jody Miller, Cindy Southworth, and participating stakeholders for the opportunity to do this work. This research was conducted on an independent, scientific basis. The work was funded in part by Abt Associates as part of a grant from the United States Department of Housing and Urban Development (HUD). The views and results expressed herein are solely those of Dr. Sweeney and do not necessarily represent those of HUD or Abt Associates.

References

1 U.S. Department of Housing and Urban Development. Homeless Management Information Systems (HMIS); Data and Technical Standards Final Notice. *Federal Register*, Vol. 69, No. 146, July 30, 2004, p. 45888-45934.

2 U.S. Department of Housing and Urban Development. *Homeless Management Information Systems (HMIS) Data and Technical Standards Final Notice; Clarification and Additional Guidance on Special Provisions for Domestic Violence Provider Shelters.* Docket No. FR 4848-N-O3. August 30, 2004.

3 U.S. Department of Housing and Urban Development. Emergency Shelter Grants Allocation History. www.hud.gov/utilities/intercept.cfm?/offices/cpd/homeless/budget/esghistory.pdf as of September 2005.

4 Northeast Ohio Coalition for the Homeless. *Overflowing Shelters: a history and recommended solutions*. April 9, 2005. www.neoch.org/what_to_do_overflowing.htm as of September 2005.

5 U.S. Conference of Mayors, A Status Report on Hunger and Homelessness in America's Cities 2001. www.usmayors.org/uscm/hungersurvey/2001/hungersurvey2001.pdf as of Sept 2005.

6 Markee, P. *Average Daily Census of Homeless Children and Adults Residing in the New York City Municipal Shelter System*. Coalition for the Homeless on behalf of New York City Department of Homeless Services and Human Resources Administration, May, 2002.

7 New York City Independent Budget Office. *Give 'Em Shelter: Various City Agencies Spend Over $900 Million on Homeless Services*. Fiscal Brief, March 2002

8 Conference Report (H.R. Report 107-272) for the Fiscal Year 2002 HUD Appropriations Act (Public Law 107-73).

9 Senate Committee Report 107-43 for the Fiscal Year 2002 HUD Appropriations Act (Public Law 107-43).

Conference Report (H.R. Report 106-988) for the Fiscal Year 2001 HUD Appropriations Act
10 (Public Law 106-377).

11 U.S. Bureau of the Census. *1990 Collection and Processing Procedures (Appendix D)* . CD-ROM Technical Documentation Project. University of Michigan. February 1998. www.lib.umich.edu/govdocs/cicdoc/cen90app/append_d.htm as of September 2005.

12 U.S. Bureau of the Census. *1996 National Survey of Homeless Assistance Providers and Clients*. Washington: 1996. www.census.gov/prod/www/nshapc/NSHAPC4.html as of September 2005.

13 M. Burt and L. Aron. *America's Homeless II: Population and Services*. Urban Institute. Washington: 2000. www.urban.org/UploadedPDF/900344_AmericasHomelessII.pdf as of Sept 2005

14 Electronic Privacy Information Center. Comments to HUD on the Matter of HMIS. Sept 2003.

15 The National Network to End Domestic Violence. Comments to HUD on the Matter of HMIS. Sept. 2003

16 National Center for Victims of Crime. Domestic Violence. As of Sept 2005, www.ncvc.org/ncvc/main.aspx?dbName=DocumentViewer&DocumentID=32347

17 U.S. Department of Justice. *Violence by Intimates: analysis of data on crimes by current or former spouses, boyfriends, and girlfriends.* NCJ-167237. March 1998.

18 S. Catania. No safe haven. *Mother Jones*, July/August 2005.

19 National HMIS TA Initiative Documents: AHAR Super Table Shells. As of Sept 2005, www.hmis.info/ta_resources_data.asp?topic_id=11

20 Privacert, Inc. *The Privacert Risk Assessment Server*. Available at www.privacert.com as of Sept 2005. Originally designed and developed by L. Sweeney.

21 Pfleeger, C. *Security in Computing*. Prentice-Hall. Upper Saddle River: 1997

22 Stinson, D. *Cryptography: Theory and Practice*. CRC Press. New York: 1995

23 Ratha, N. and Bolle, R. Automatic Fingerprint Recognition Systems. Springer-Verlag. New York: 2004

24 Russell, R. Soundex. U.S. Patent 1,261,167 April 2, 1918.

25 Record Linkage Techniques -- 1997: Proceedings of an International Workshop and Exposition. National Research Council, 1999.

26 Sweeney, L. *Inconsistent Hashing and the Notion of Single-Use Identifying Numbers.* Carnegie Mellon University, School of Computer Science, Data Privacy Lab White Paper Series LIDAP-WP13. Pittsburgh, PA: 2005.

27 Edo-Eket, S. and Sweeney, L. *Detecting Bio-Terrorist Attacks and Naturally Occurring Outbreaks Over a Distributed Network While Protecting Privacy and Confidentiality: the PrivaSum Protocol.* Carnegie Mellon University, School of Computer Science, Technical Report CMU-ISRI-04-111.

28 J. Benaloh and M. de Mare. One-way accumulators: a decentralized alternative to digital signatures. In *Proceedings of Advances in Cryptology - EUROCRYPT '93, Lecture Notes in Computer Science*, v 765, pages 274-285, Lofthus, Norway, 1994.

29 Sweeney, L. and Shamos, M. *A Multiparty Computation for Randomly Ordering Players and Making Random Selections*. Carnegie Mellon University, School of Computer Science, Technical Report, CMU-ISRI-04-126. Pittsburgh: July 2004. privacy.cs.cmu.edu/dataprivacy/projects/randomorder/index.html

30 Sweeney, L. k-anonymity: a model for protecting privacy. I*nternational Journal on Uncertainty, Fuzziness and Knowledge-based Systems*, 10 (5), 2002; 557-570. privacy.cs.cmu.edu/people/sweeney/kanonymity.html.

31 Sweeney, L. Achieving k-anonymity privacy protection using generalization and suppression. *International Journal on Uncertainty, Fuzziness and Knowledge-based Systems*, 10 (5), 2002; 571-588. privacy.cs.cmu.edu/people/sweeney/kanonymity2.html.

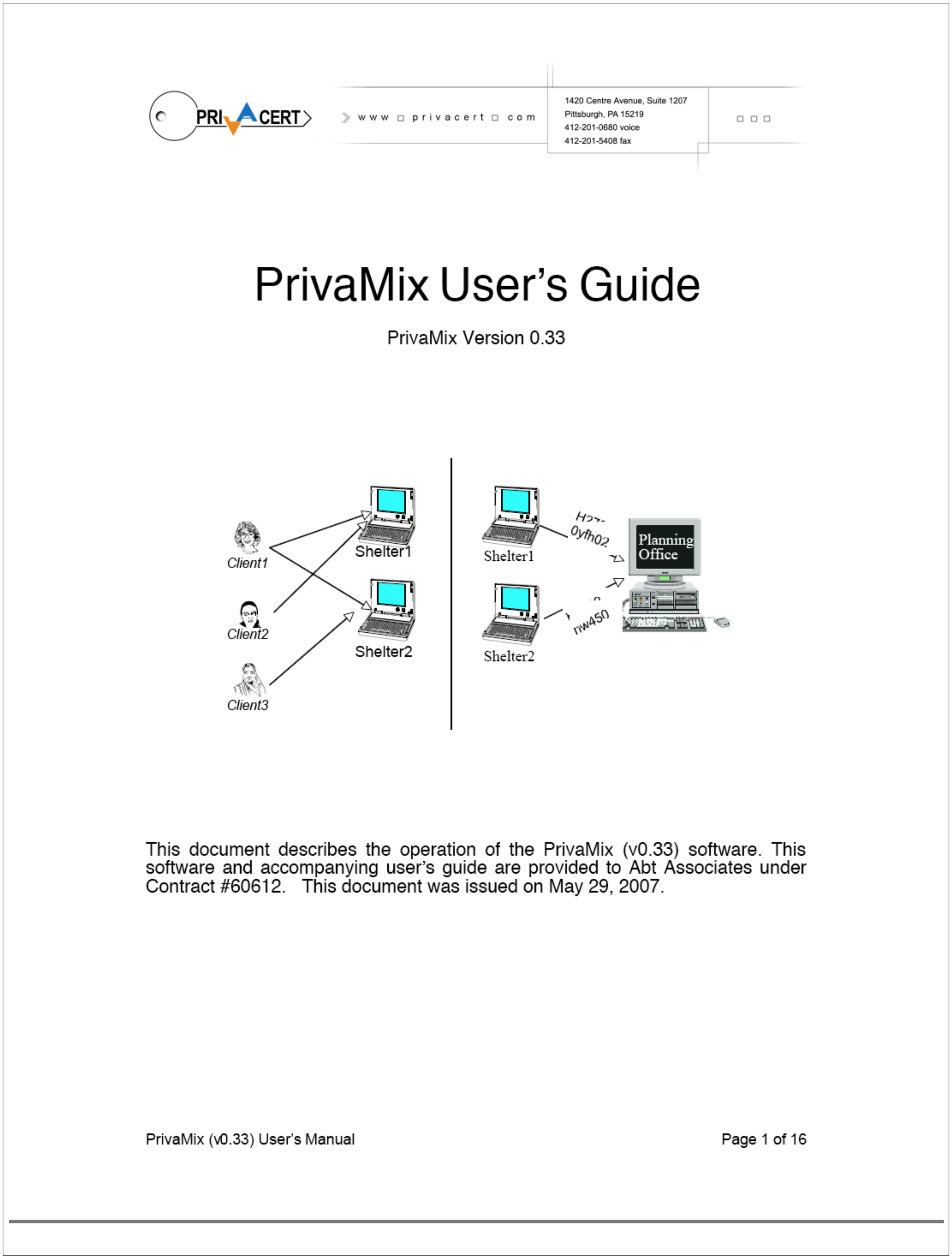
32 Sweeney, L. The Search for a P3Tracker Hash Function: a research notebook. Carnegie Mellon University. LIDAP Working Paper 31. Pittsburgh: April 2006.

33 Russell, R. *Soundex*. U.S. Patent 1261167. April 2, 1918.

34 Sokol, B. *PrivaMix Proof-of-Concept Report*. Abt Associates. July 2007.

Appendix A

PrivaMix User's Guide



PRIVACERT

www.privacert.com

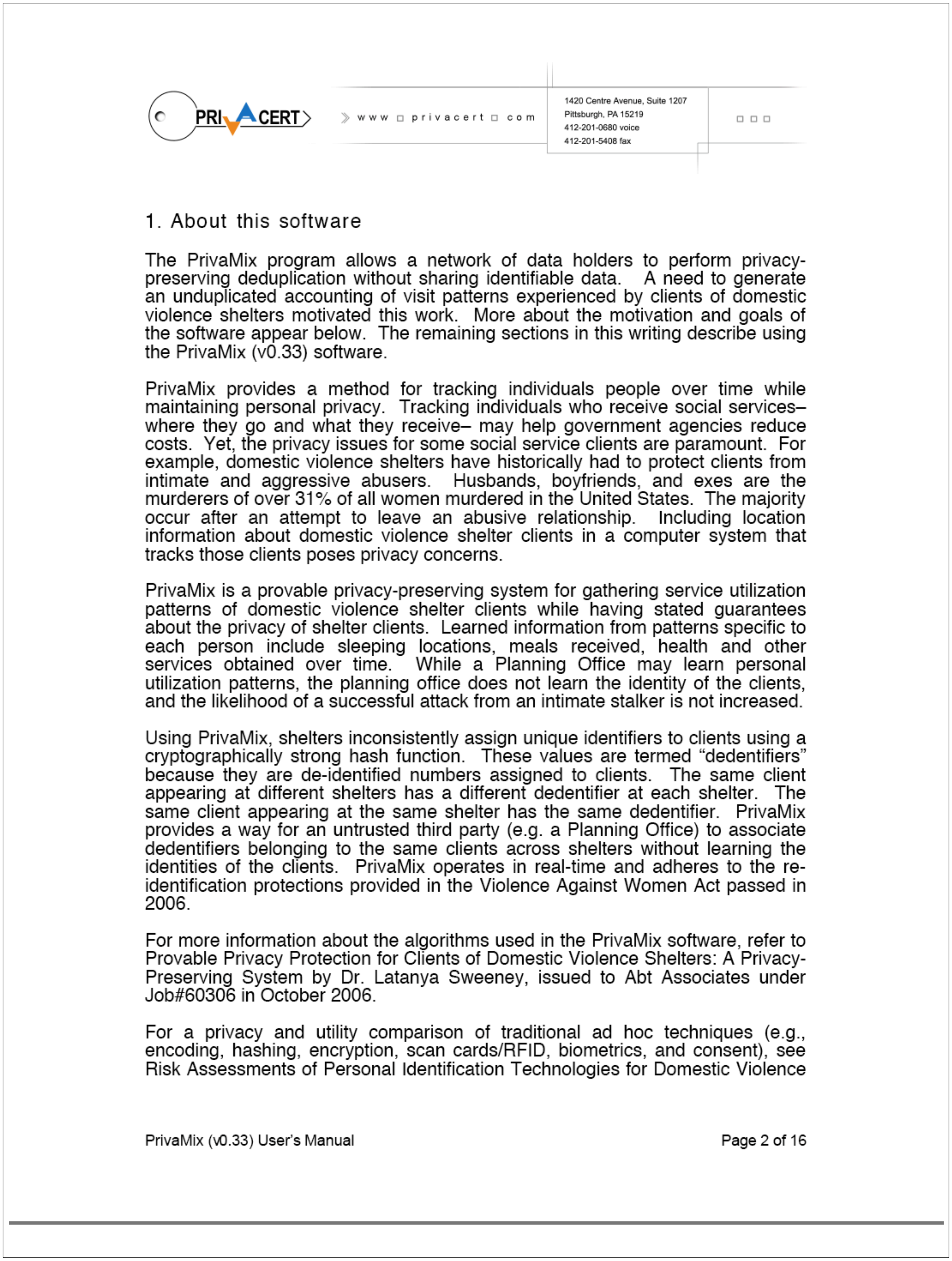
1420 Centre Avenue, Suite 1207
Pittsburgh, PA 15219
412-201-0680 voice
412-201-5408 fax

PrivaMix User's Guide

PrivaMix Version 0.33

This document describes the operation of the PrivaMix (v0.33) software. This software and accompanying user's guide are provided to Abt Associates under Contract #60612. This document was issued on May 29, 2007.

PrivaMix (v0.33) User's Manual Page 1 of 16



PRIVACERT

www.privacert.com

1420 Centre Avenue, Suite 1207
Pittsburgh, PA 15219
412-201-0680 voice
412-201-5408 fax

1. About this software

The PrivaMix program allows a network of data holders to perform privacy-preserving deduplication without sharing identifiable data. A need to generate an unduplicated accounting of visit patterns experienced by clients of domestic violence shelters motivated this work. More about the motivation and goals of the software appear below. The remaining sections in this writing describe using the PrivaMix (v0.33) software.

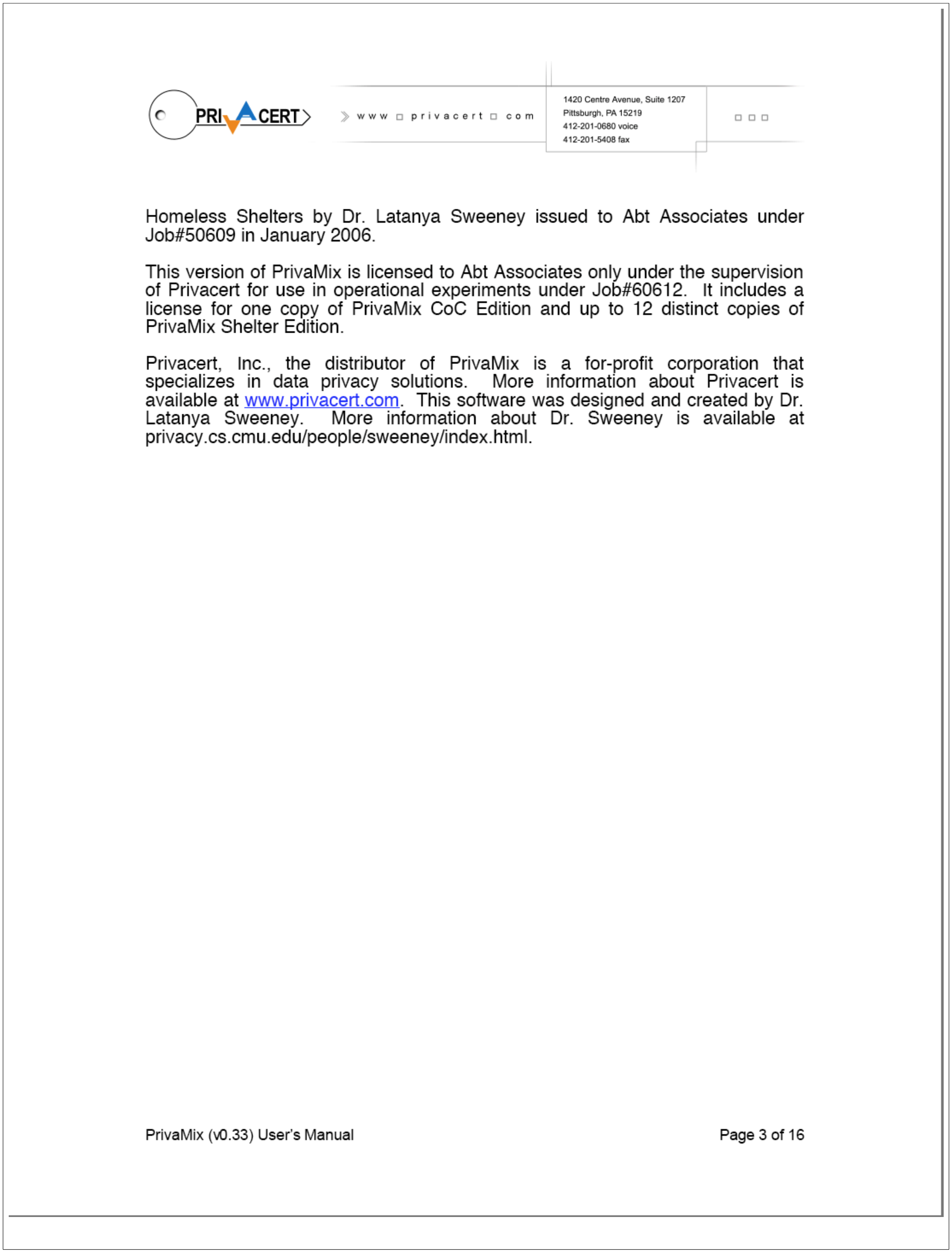
PrivaMix provides a method for tracking individuals people over time while maintaining personal privacy. Tracking individuals who receive social services– where they go and what they receive– may help government agencies reduce costs. Yet, the privacy issues for some social service clients are paramount. For example, domestic violence shelters have historically had to protect clients from intimate and aggressive abusers. Husbands, boyfriends, and exes are the murderers of over 31% of all women murdered in the United States. The majority occur after an attempt to leave an abusive relationship. Including location information about domestic violence shelter clients in a computer system that tracks those clients poses privacy concerns.

PrivaMix is a provable privacy-preserving system for gathering service utilization patterns of domestic violence shelter clients while having stated guarantees about the privacy of shelter clients. Learned information from patterns specific to each person include sleeping locations, meals received, health and other services obtained over time. While a Planning Office may learn personal utilization patterns, the planning office does not learn the identity of the clients, and the likelihood of a successful attack from an intimate stalker is not increased.

Using PrivaMix, shelters inconsistently assign unique identifiers to clients using a cryptographically strong hash function. These values are termed "dedentifiers" because they are de-identified numbers assigned to clients. The same client appearing at different shelters has a different dedentifier at each shelter. The same client appearing at the same shelter has the same dedentifier. PrivaMix provides a way for an untrusted third party (e.g. a Planning Office) to associate dedentifiers belonging to the same clients across shelters without learning the identities of the clients. PrivaMix operates in real-time and adheres to the re-identification protections provided in the Violence Against Women Act passed in 2006.

For more information about the algorithms used in the PrivaMix software, refer to Provable Privacy Protection for Clients of Domestic Violence Shelters: A Privacy-Preserving System by Dr. Latanya Sweeney, issued to Abt Associates under Job#60306 in October 2006.

For a privacy and utility comparison of traditional ad hoc techniques (e.g., encoding, hashing, encryption, scan cards/RFID, biometrics, and consent), see Risk Assessments of Personal Identification Technologies for Domestic Violence



PRIVACERT

www.privacert.com

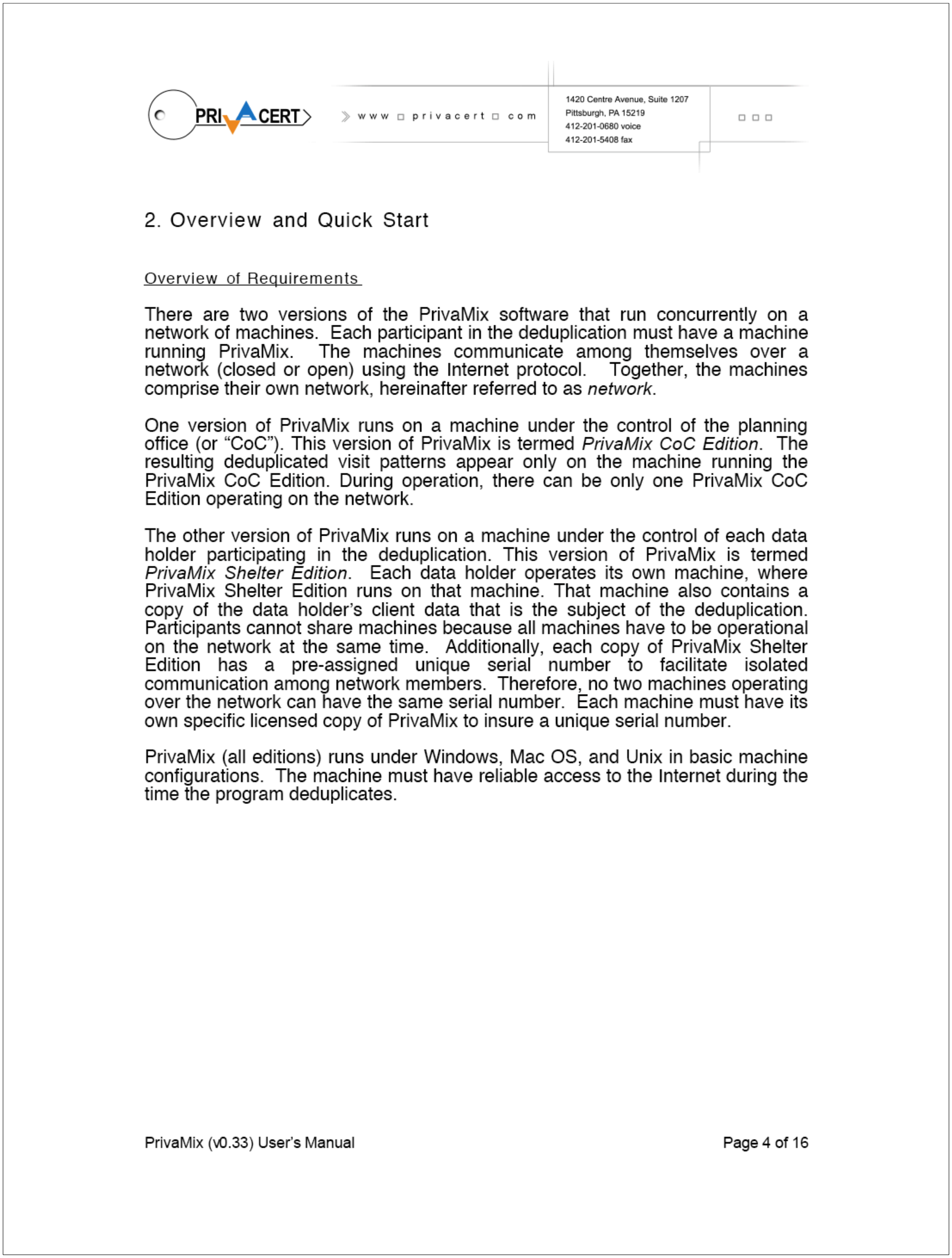
1420 Centre Avenue, Suite 1207
Pittsburgh, PA 15219
412-201-0680 voice
412-201-5408 fax

Homeless Shelters by Dr. Latanya Sweeney issued to Abt Associates under Job#50609 in January 2006.

This version of PrivaMix is licensed to Abt Associates only under the supervision of Privacert for use in operational experiments under Job#60612. It includes a license for one copy of PrivaMix CoC Edition and up to 12 distinct copies of PrivaMix Shelter Edition.

Privacert, Inc., the distributor of PrivaMix is a for-profit corporation that specializes in data privacy solutions. More information about Privacert is available at www.privacert.com. This software was designed and created by Dr. Latanya Sweeney. More information about Dr. Sweeney is available at privacy.cs.cmu.edu/people/sweeney/index.html.

PrivaMix (v0.33) User's Manual Page 3 of 16



2. Overview and Quick Start

Overview of Requirements

There are two versions of the PrivaMix software that run concurrently on a network of machines. Each participant in the deduplication must have a machine running PrivaMix. The machines communicate among themselves over a network (closed or open) using the Internet protocol. Together, the machines comprise their own network, hereinafter referred to as *network*.

One version of PrivaMix runs on a machine under the control of the planning office (or "CoC"). This version of PrivaMix is termed *PrivaMix CoC Edition*. The resulting deduplicated visit patterns appear only on the machine running the PrivaMix CoC Edition. During operation, there can be only one PrivaMix CoC Edition operating on the network.

The other version of PrivaMix runs on a machine under the control of each data holder participating in the deduplication. This version of PrivaMix is termed *PrivaMix Shelter Edition*. Each data holder operates its own machine, where PrivaMix Shelter Edition runs on that machine. That machine also contains a copy of the data holder's client data that is the subject of the deduplication. Participants cannot share machines because all machines have to be operational on the network at the same time. Additionally, each copy of PrivaMix Shelter Edition has a pre-assigned unique serial number to facilitate isolated communication among network members. Therefore, no two machines operating over the network can have the same serial number. Each machine must have its own specific licensed copy of PrivaMix to insure a unique serial number.

PrivaMix (all editions) runs under Windows, Mac OS, and Unix in basic machine configurations. The machine must have reliable access to the Internet during the time the program deduplicates.

Quick Start

If the network has been pre-configured, all the default settings will allow immediate operation. In this case, participants do the following:

1. *Data holders*: Copy your client data file to c:\data.csv. The file must have this name and be located in this directory.
2. *CoC*: Load the program and then click on the *Deduplicate* button. See Figure 1(a) below.
3. *Data holders*: Load the program and then click on the *Deduplicate* button. See Figure 1(b) below.
4. *CoC*: The results of the deduplication appear in the files c:\results1.csv and results2.csv when the button is renamed to *Exit*.

If your network or machine has not been pre-configured or the default settings need to be confirmed, checked, or modified, see the remaining sections of this writing for specifics on how to make changes and check files.

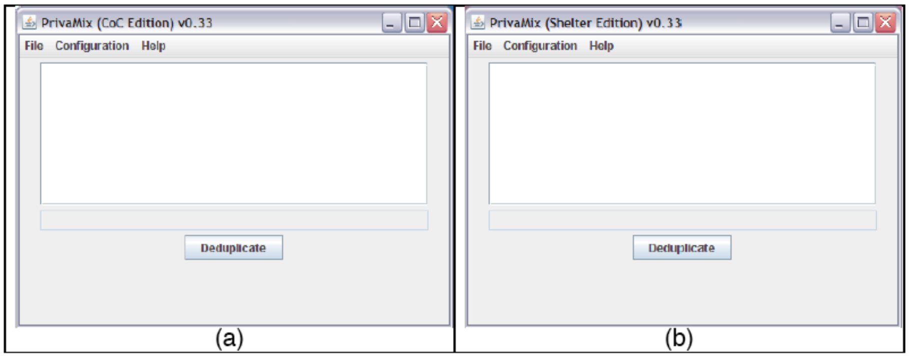


Figure 1. Program screen for CoC Edition (a) and for Shelter Edition (b). Click on the "Deduplicate" button to run the program.

3. Shelter Edition

Below is a description of the basic menu commands in PrivaMix Shelter Edition. The order in which a command appears in this writing is organized around the program display –File, Configuration, and Help. Click on the *Deduplicate* button to execute the program.

Deduplicate

Click the *Deduplicate* button, which prominently appears on the main program screen to execute the program. If a problem occurs, use the menu commands to change the configuration to match the selected data file and network. Figure 2 shows the Deduplicate button as it appears in the program window.

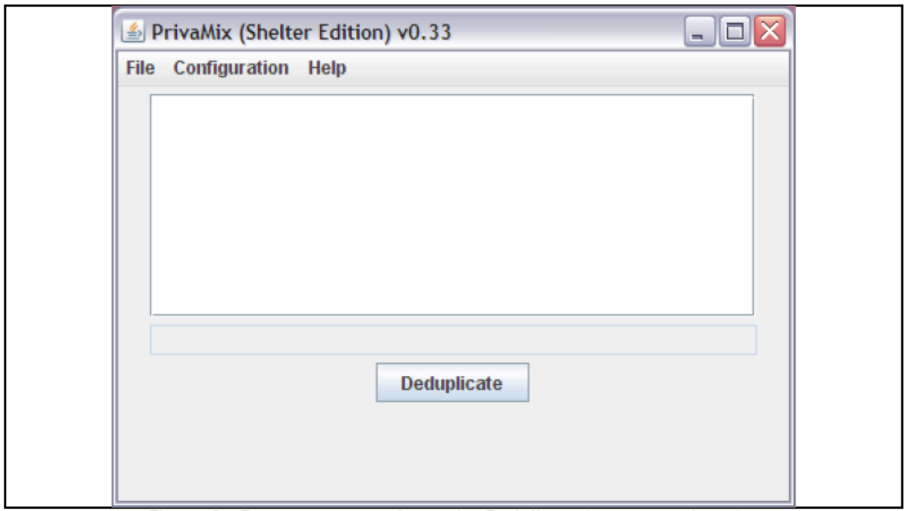


Figure 2. Program screen shows the Dedplicate button prominently.

File Menu

The *File* Menu has operations related to the file that contains the client information that is the subject of the deduplication. Figure 3 shows the available operations, which include *Select* and *Check*. These operations are described below. The File Menu also includes an *Exit* operation to terminate the program.

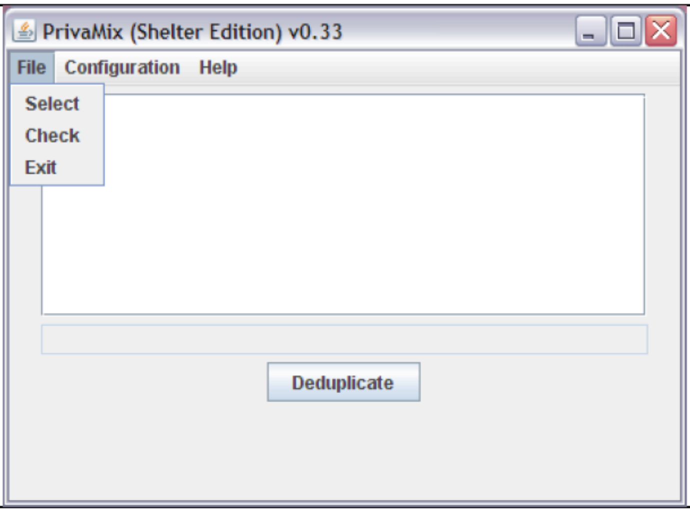


Figure 3. File Menu for Shelter Edition.

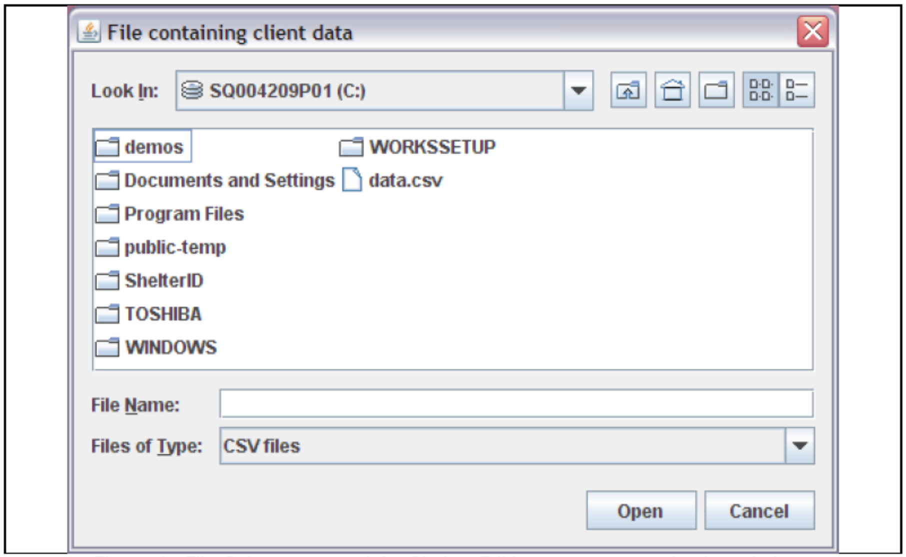
File-Select

Clicking on *File*, then *Select* displays a window for navigating through the file system in order to locate the file that contains the client information that is the subject of the deduplication. This file should be a CSV (comma-delimited file), whose filename ends in ".txt", ".TXT", ".csv", or ".CSV". Figure 4 shows an example of the navigation window that appears.

To see which data file is currently selected, use the Configuration-Current filename command.

To check the integrity of the format file to make sure it matches the program's current expectations, use the File-Check command.

To run the program, click the Deduplicate button.



File containing client data

Look In: SQ004209P01 (C:)

File Name:

Files of Type: CSV files

Figure 4. File-Select command for Shelter Edition pops up a navigation window.

File-Check

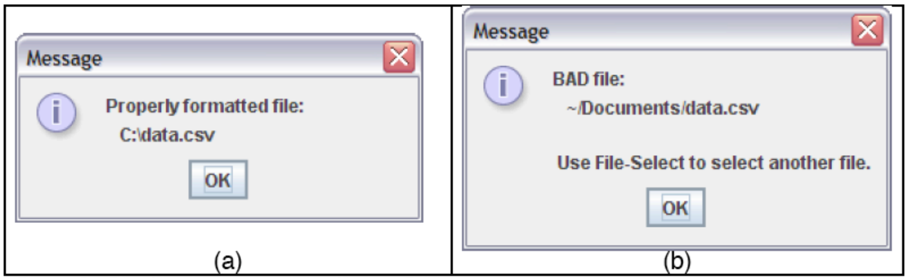
Clicking on *File*, then *Check* runs an integrity check on the data file currently selected and reports the result as good or bad. Figure 5 shows the good and bad result messages. The filename of the file that is checked appears in the pop-up window, as shown in Figure 5.

To see what data file is already selected, use the Configuration-Current filename command.

To change the data file to use, use the File-Select command.

To change the format expected of the file, use the Configuration-CSV format command.

To run the program, click the Deduplicate button.



Properly formatted file:
C:\data.csv

BAD file:
~/Documents/data.csv

Use File-Select to select another file.

Figure 5. File-Check command for Shelter Edition pops up a message about the integrity of the file as good (a) or bad (b).

Configuration Menu

The *Configuration* menu contains commands for altering default values used throughout the system. Figure 6 shows a list of available configuration commands.

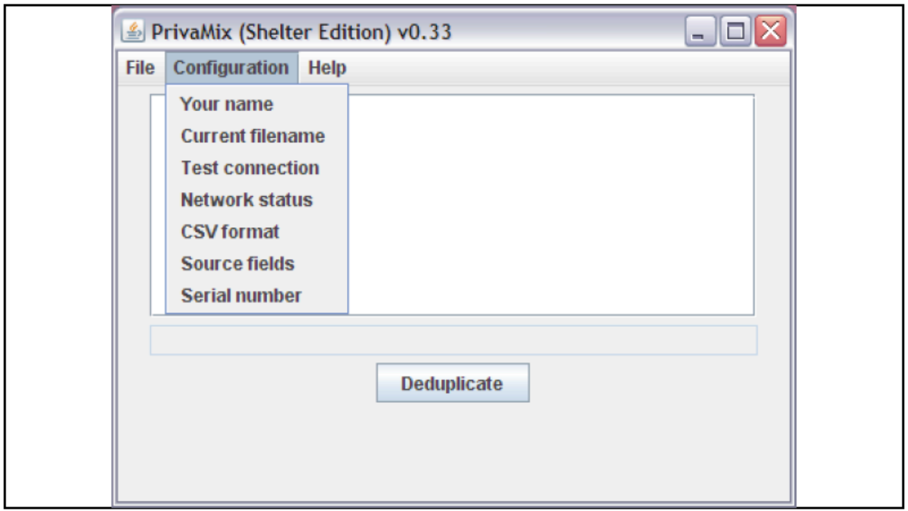


Figure 6. Configuration menu for Shelter Edition.

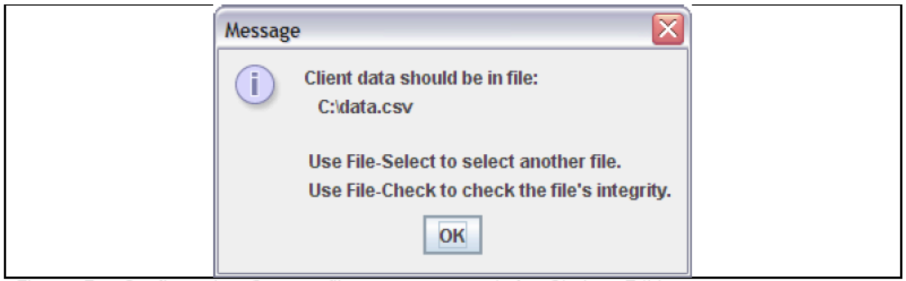
Configuration-Current filename

Clicking on *Configuration*, then *Current filename* displays the name of the file currently selected as the file containing the client information for deduplication. Figure 7 shows an example of the pop-up window.

To change the data file to use, use the File-Select command.

To change the format expected of the file, use the Configuration-CSV format command.

To run the program, click the Deduplicate button.



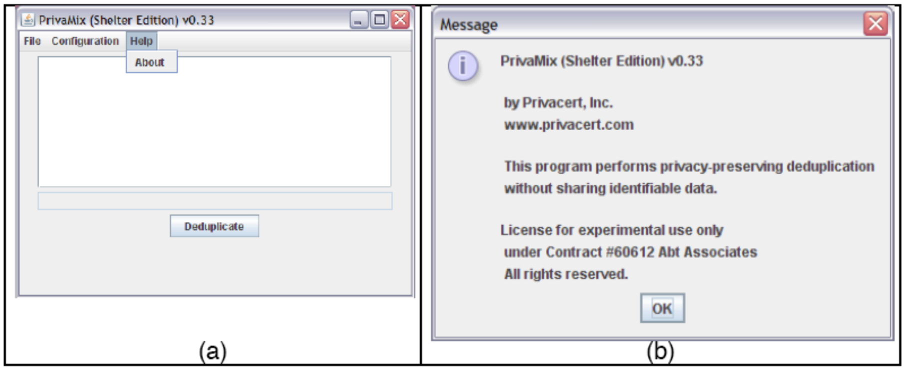
Client data should be in file:
C:\data.csv

Use File-Select to select another file.
Use File-Check to check the file's integrity.

Figure 7. Configuration-Current filename command for Shelter Edition pops up a message displaying the filename of the file currently selected for processing.

Help Menu and the Help-About Command

The *Help* menu contains the About command which displays version and license information for the copy of PrivaMix running. Figure 8 (a) shows the Help menu and Figure 8(b) shows the pop-up window that appears with the Help-About command.



PrivaMix (Shelter Edition) v0.33

by Privacert, Inc.
www.privacert.com

This program performs privacy-preserving deduplication without sharing identifiable data.

License for experimental use only
under Contract #60612 Abt Associates
All rights reserved.

(a) (b)

Figure 8. Help menu (a) and the results of the Help-About command (b).

3. CoC Edition

Below is a description of the basic menu commands in PrivaMix CoC Edition. These commands are similar to those found in the Shelter Edition, mentioned previously. Click on the Dediplicate program to execute the program.

Deduplicate

Click the *Deduplicate* button, which prominently appears on the main program screen to execute the program. If a problem occurs, use the menu commands to change the configuration to match the selected data file and network. Figure 9 shows the Deduplicate button as it appears in the program window.

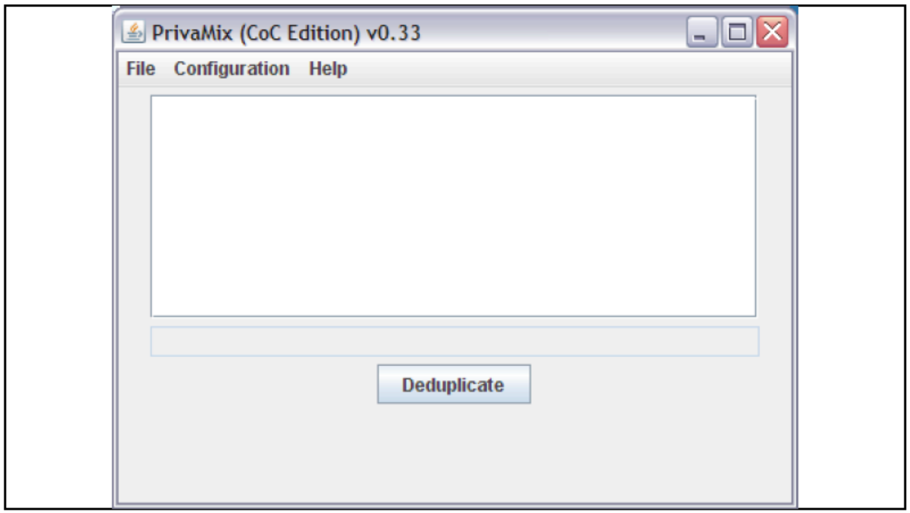


Figure 9. Program screen shows the Dedplicate button prominently.

File Menu (CoC Edition)

The *File* Menu has the *Save as* operation, which allows the user to provide the names of the files files that will store the results of the deduplication. The File Menu also includes an *Exit* operation to terminate the program. Figure 10 shows the available operations.

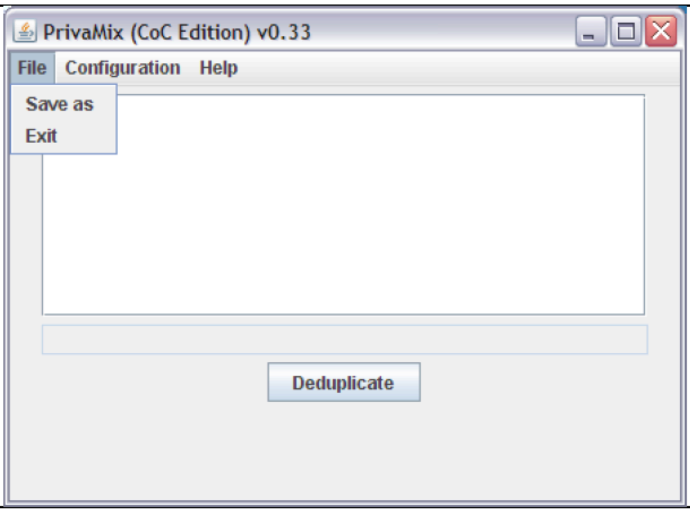


Figure 10. File Menu for CoC Edition.

Configuration Menu (CoC Edition)

The *Configuration* menu contains commands for altering default values used throughout the system. Figure 11 shows a list of available configuration commands.

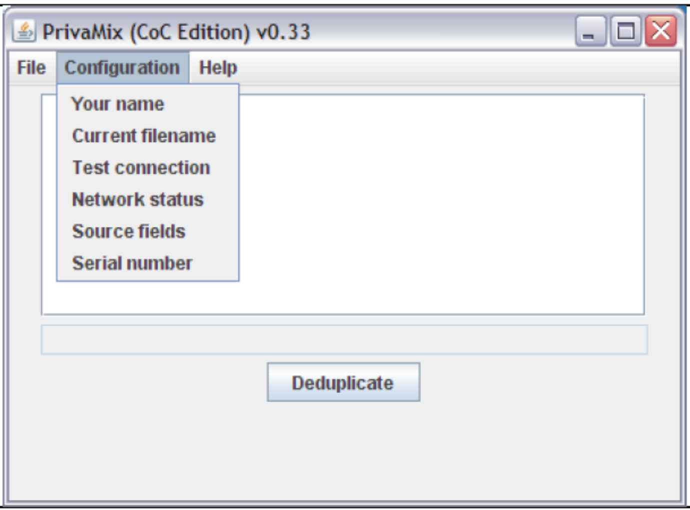
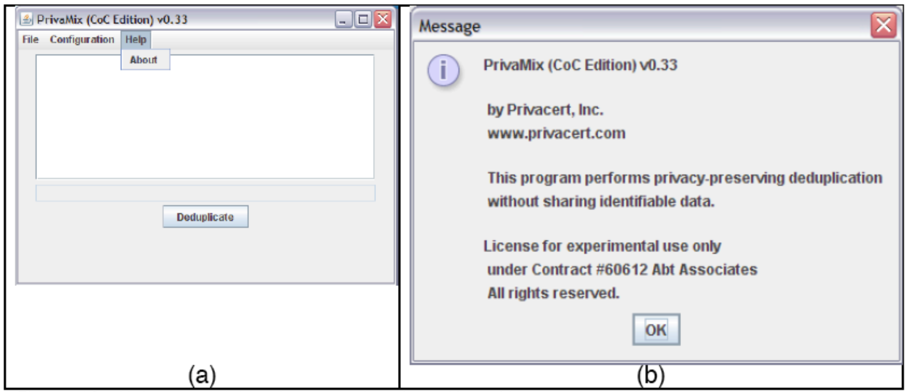


Figure 11. Configuration menu for CoC Edition.

Help Menu and the Help-About Command (CoC Edition)

The *Help* menu contains the *About* command which displays version and license information for the copy of PrivaMix running. Figure 12 (a) shows the Help menu and Figure 12(b) shows the pop-up window that appears with the Help-About command.



PrivaMix (CoC Edition) v0.33

by Privacert, Inc.
www.privacert.com

This program performs privacy-preserving deduplication without sharing identifiable data.

License for experimental use only
under Contract #60612 Abt Associates
All rights reserved.

Figure 12. Help menu (a) and the results of the Help-About command (b).

Index