An Overview of the NSF-ITR-Census Bureau Synthetic Data Project

John M. Abowd
Cornell University
November 21, 2008
3rd IAB Workshop on Confidentiality and Disclosure - SDC for Microdata

Acknowledgements and Disclaimer

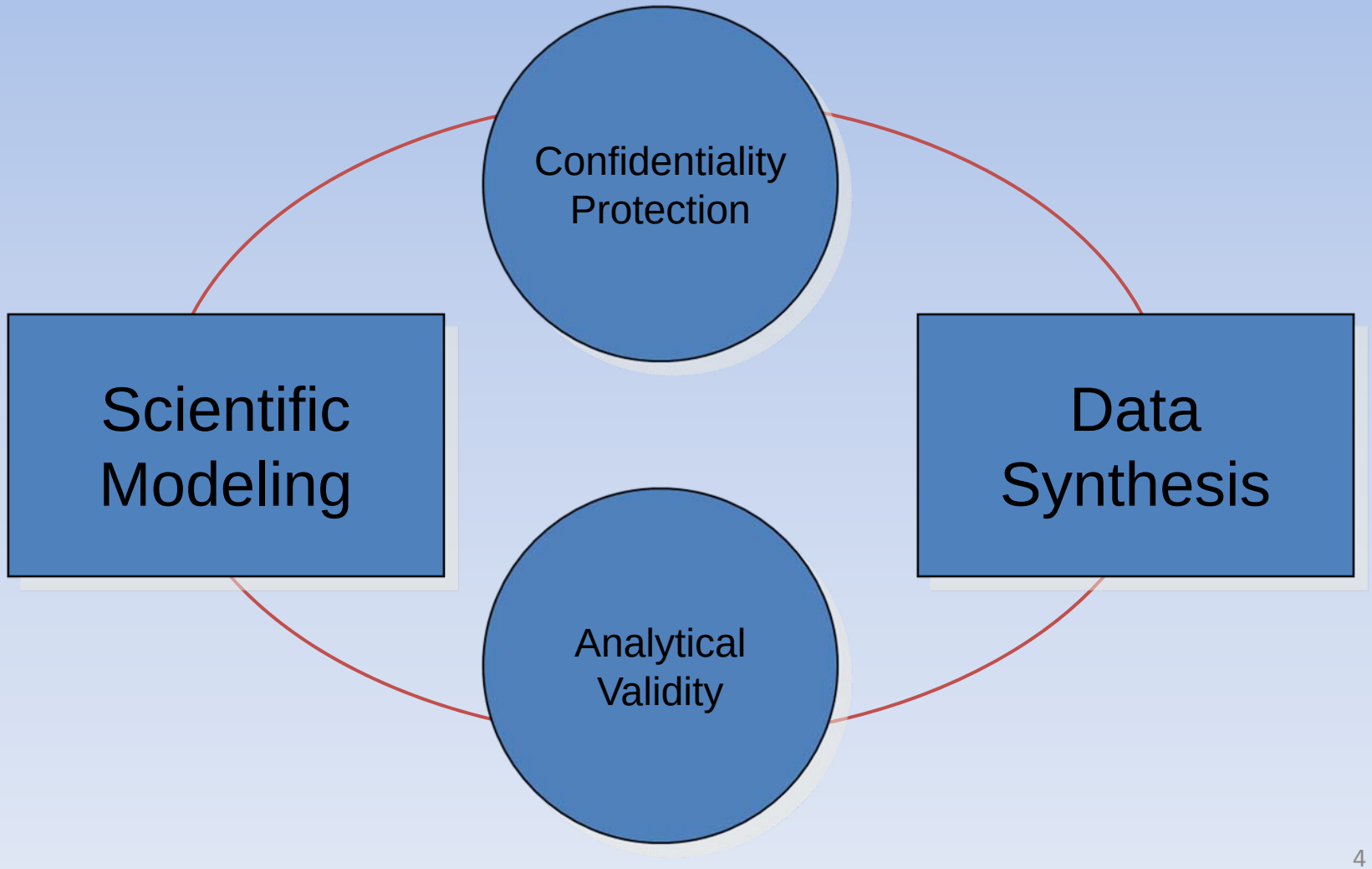
- This work was partially supported by the National Science Foundation (grants SES-0339191, SES-0427889 and CNS-0627680) and the United States Census Bureau.

- All statistical materials in this presentation have been reviewed for disclosure avoidance.
- The opinions are those of the author and not the National Science Foundation nor the Census Bureau.

Outline

- Foundations of the NSF-ITR-Census Bureau project
- Criteria for “good” synthetic data
- Core projects
- The Cornell Virtual Research Data Center

The Research – Synthetic Data Feedback Cycle



Definition of Synthetic Data

$$X \equiv \text{confidential data}$$

$$\Pr\left[\tilde{X} \middle| X\right] \equiv \text{PPD of } \tilde{X} \text{ given } X$$

$$\text{Release data are samples of } \tilde{X}$$

- Synthetic data are created by estimating the posterior predictive distribution (PPD) of the release data given the confidential data; then sampling release data from the PPD conditioning on the actual confidential values.
- The PPD is a parameter-free forecasting model for new values of the complete data matrix that conditions on all values of the underlying confidential data.

Connection to Randomized Sanitizers

$$X \equiv \text{confidential data}$$

$$U \equiv \text{random noise}$$

$$\text{San}(X, U): (X, U) \rightarrow \tilde{X}$$

$$\Pr\left[\tilde{X} \middle| X\right] \equiv \text{probability of } \tilde{X} \text{ given } X$$

- A randomized sanitizer creates a conditional probability distribution for the release data given the confidential data.
- The randomness in a sanitizer is induced by the properties of the distribution of *U*.
- The PPD is just a particular randomized sanitizer.

Disclosure Limitation Definitions

$$X = x^{(1)} \text{ and } X = x^{(2)}$$

$\tilde{X} = \tilde{x}$, realization of the synthesizer

- Consider two confidential data matrices that differ in only a single row, $x^{(1)}$ and $x^{(2)}$.
- Use the PPD to evaluate the probability of a particular release data set given the two different confidential data sets.

Synthetic Data Can Leak Information about a Single Entity

$$\Pr\left[\tilde{X} = \tilde{x} \middle| X = x^{(1)}\right] \neq \Pr\left[\tilde{X} = \tilde{x} \middle| X = x^{(2)}\right]$$

- Changing a single row of the confidential data matrix changes the PPD or the random sanitizer.
- The PPD or the random sanitizer define the transition probabilities from the confidential data to the release data.
- True for all SDL procedures that infuse noise.

Connection Between Synthetic Data and Differential Privacy

$$\frac{\dfrac{\Pr\left[X = x^{(1)} \middle| \tilde{X} = \tilde{x}\right]}{\Pr\left[X = x^{(2)} \middle| \tilde{X} = \tilde{x}\right]}}{\dfrac{\Pr[X = x^{(1)}]}{\Pr[X = x^{(2)}]}} = \frac{\Pr\left[\tilde{X} = \tilde{x} \middle| X = x^{(1)}\right]}{\Pr[\tilde{X} = \tilde{x} | X = x^{(2)}]}$$

The posterior odds ratio for the gain in information about a single row of X is equal to the differential privacy from the randomized sanitizer that creates release data by sampling from the specified conditional distribution.

Connection Between Differential Privacy and Inferential Disclosure

$$\frac{\frac{\Pr\left[X = x^{(1)} \middle| \tilde{X} = \tilde{x}\right]}{\Pr\left[X = x^{(2)} \middle| \tilde{X} = \tilde{x}\right]}}{\frac{\Pr[X = x^{(1)}]}{\Pr[X = x^{(2)}]}} = \frac{\Pr\left[\tilde{X} = \tilde{x} \middle| X = x^{(1)}\right]}{\Pr[\tilde{X} = \tilde{x} | X = x^{(2)}]}$$

The posterior odds ratio for the gain in information about a single row of X is the Dalenius (1977) definition of an inferential disclosure. Bounding the differential privacy therefore bounds the inferential disclosure.

Goals of Synthetic Data

- Analytical validity
  - Statistical inferences based on the synthetic data should be “similar” to those based on the underlying confidential data.
- Confidentiality protection
  - The perturbation of the confidential data induced by replacing some, or all, of the values with draws from the PPD should be adequate to “protect” the confidential data.

Formal Models of Analytical Validity

- Unconditional analytical validity
  - The synthetic data process delivers the same inferences as the process that generated the confidential data. This property depends on both the synthesizer and the design of the confidential data
- Conditional analytical validity
  - The synthetic data process delivers the same inferences as the realized confidential data
- The Rubin (1993) inference validity was based on using multiple samples (implicates) from the PPD. It is an unconditional analytical validity model.

Project 1: Survey of Income and Program Participation Synthetic Beta

- Based on 1990-1993 and 1996 SIPP panels
- Linked to complete earnings and benefit histories from SSA (1950 to 2003)
- 16-implicate partially synthetic file with 633 variables released as Beta in November 2007
- Any beta user can have the analysis performed on the underlying confidential data by following the procedure outlined by the Census Bureau

Survey of Income and Program Participation (SIPP) - Mozilla Firefox

File Edit View History Bookmarks Tools Help

http://www.census.gov/sipp/index.html

sipp synthetic beta

Most Visited Latest Headlines Gmail ACES Econ 1110 NYT WSJ RCP VirtualRDCWiki CUINFO John M. Abowd cycling

U.S. Census Bureau

Survey of Income and Program Participation

SIPP

- Introduction to SIPP
- SIPP Survey Content
- Technical Information
- Using & Linking Files
- SIPP Publications
- Access SIPP Data
- Access SIPP Synthetic Data
- SIPP Small Grants Awards
- 2004 Panel Release Schedule

- User Notes/ ListServe/News
- SIPP Users' Guide
- SIPP Tutorial
- Technical Documentation
- SIPP Help
- re-engineered SIPP
- Contact re-engineered SIPP

(Formerly, DEWS)

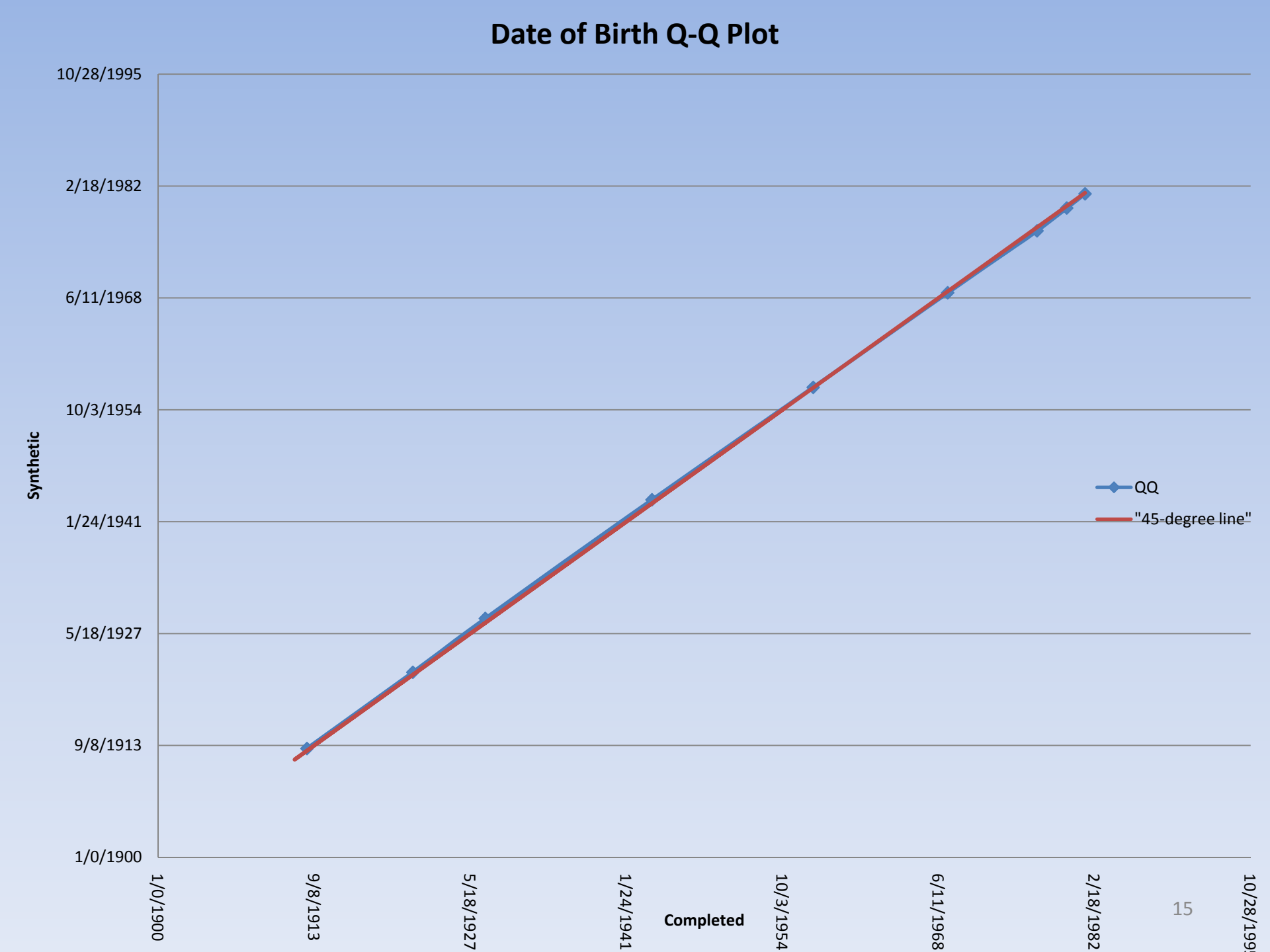
URL: http://www.census.gov/sipp/

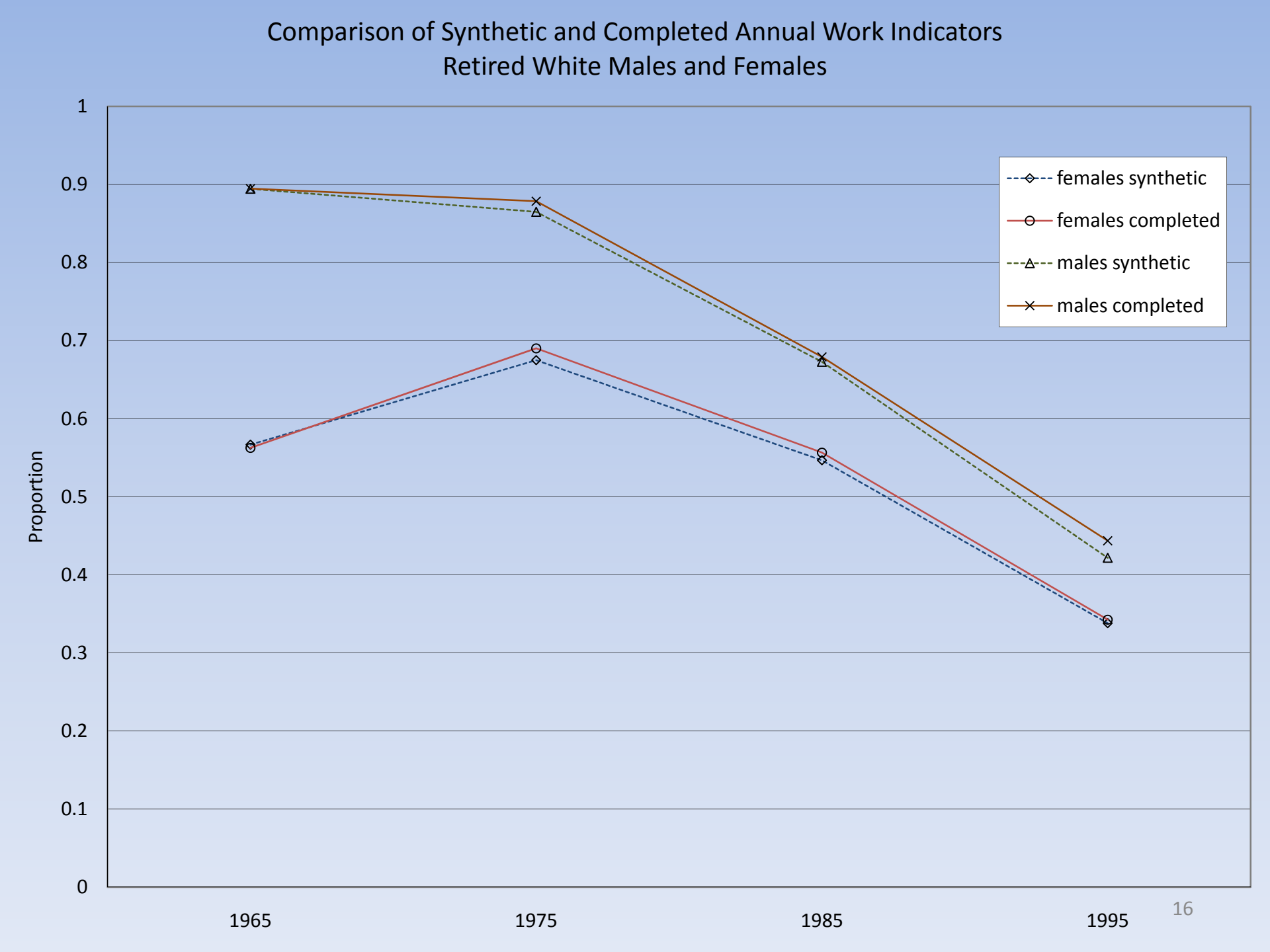
Source: U.S. Census Bureau, Demographics Survey Division, Survey of Income and Program Participation branch
Created: February 14, 2002
Last revised: January 2, 2008

Census Bureau Links: Home • Search • Subjects A-Z • FAQs • Data Tools • Catalog • Census 2000 • Quality • Privacy Policy • Contact Us

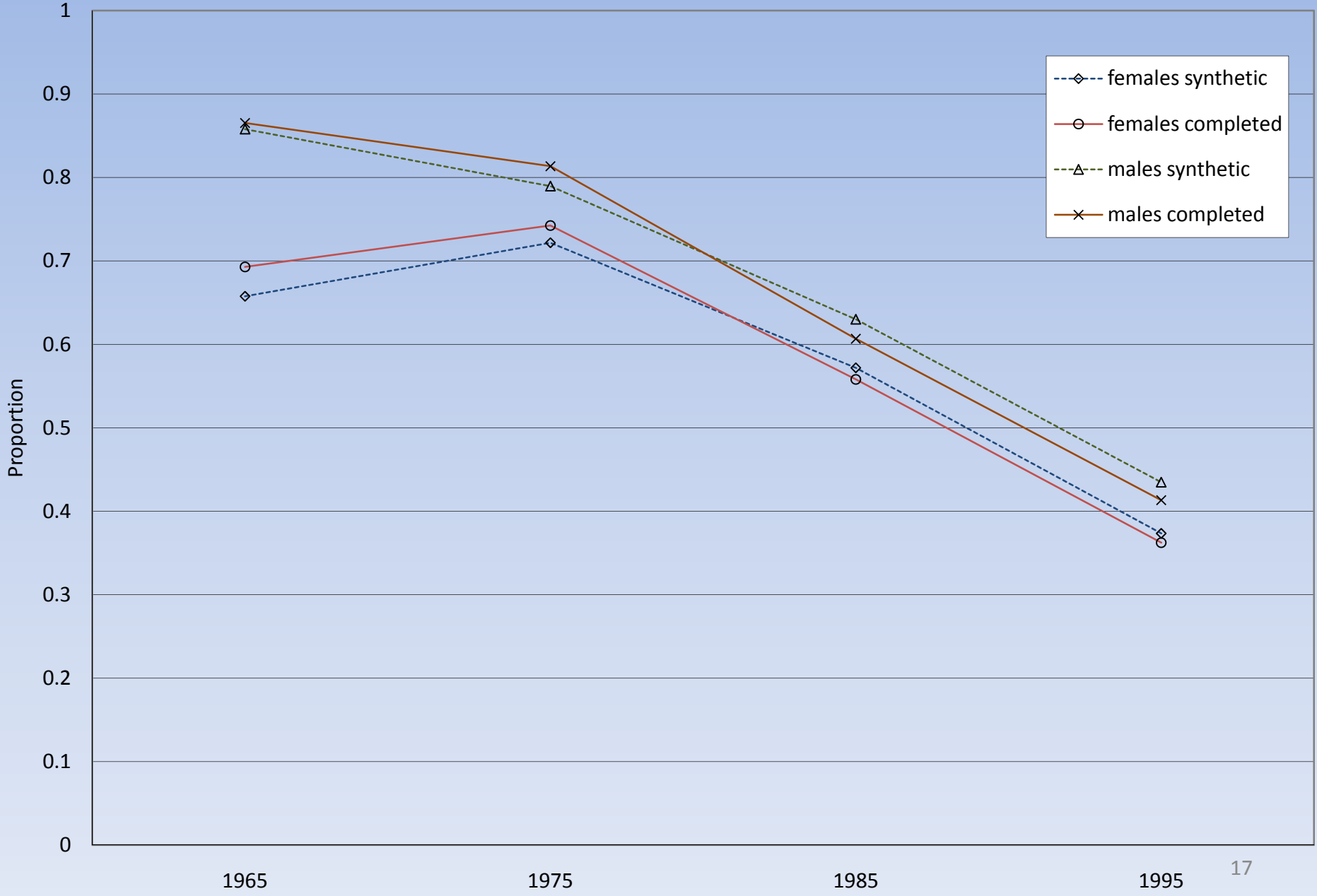
USCENSUSBUREAU
Helping You Make Informed Decisions

Done


Date of Birth Q-Q Plot
10/28/1995
2/18/1982
6/11/1968
10/3/1954
1/24/1941
5/18/1927
9/8/1913
1/0/1900
Synthetic
1/0/1900
9/8/1913
5/18/1927
1/24/1941
10/3/1954
6/11/1968
2/18/1982
10/28/1995
Completed
QQ
"45-degree line"

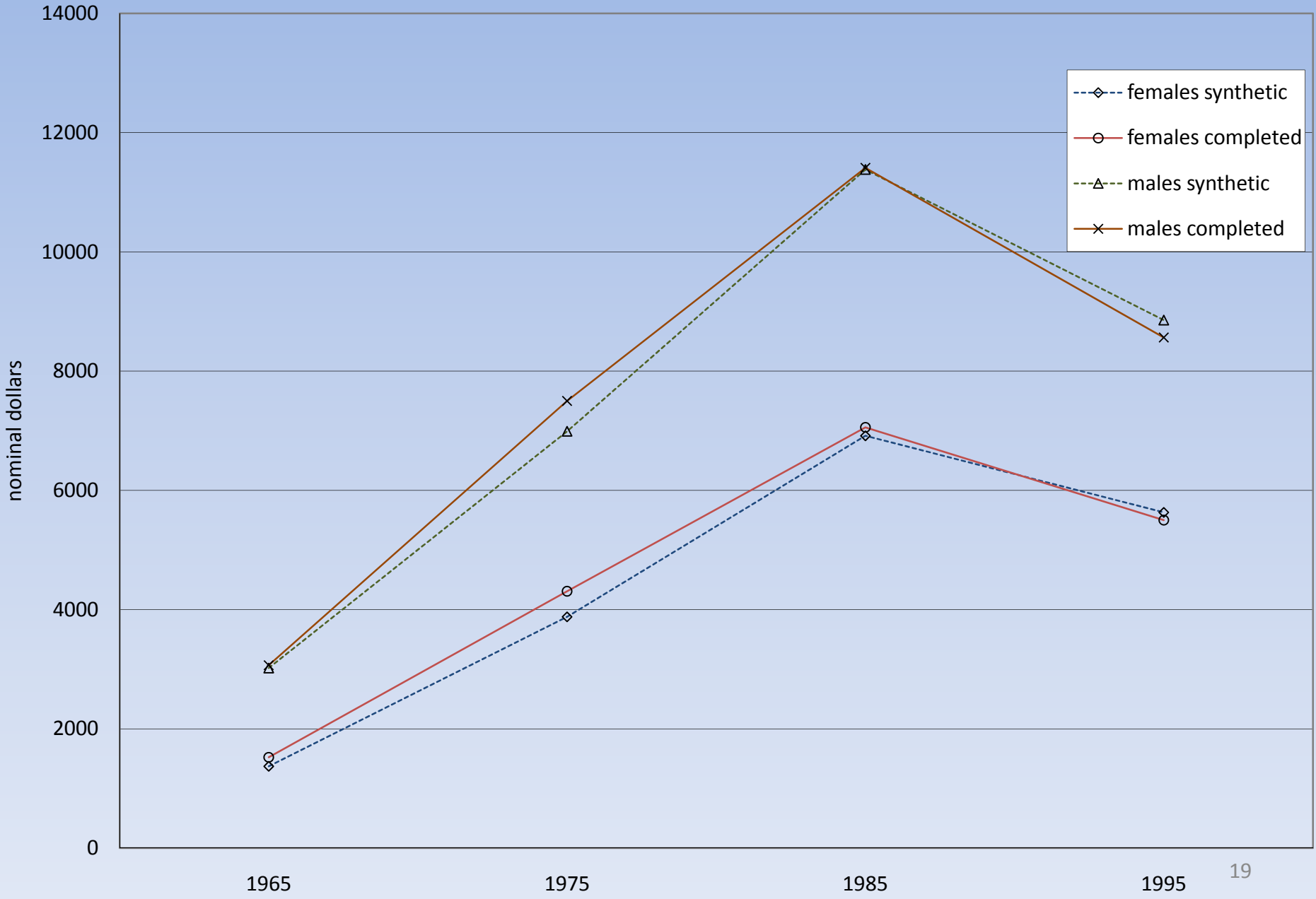

Comparison of Synthetic and Completed Annual Work Indicators
Retired White Males and Females
Proportion
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
1965
1975
1985
1995
females synthetic
females completed
males synthetic
males completed

Comparison of Synthetic and Completed Annual Work Indicators
Retired Black Males and Females



Comparison of Synthetic and Completed Earnings
Retired White Males and Females

Comparison of Synthetic and Completed Earnings
Retired Black Males and Females



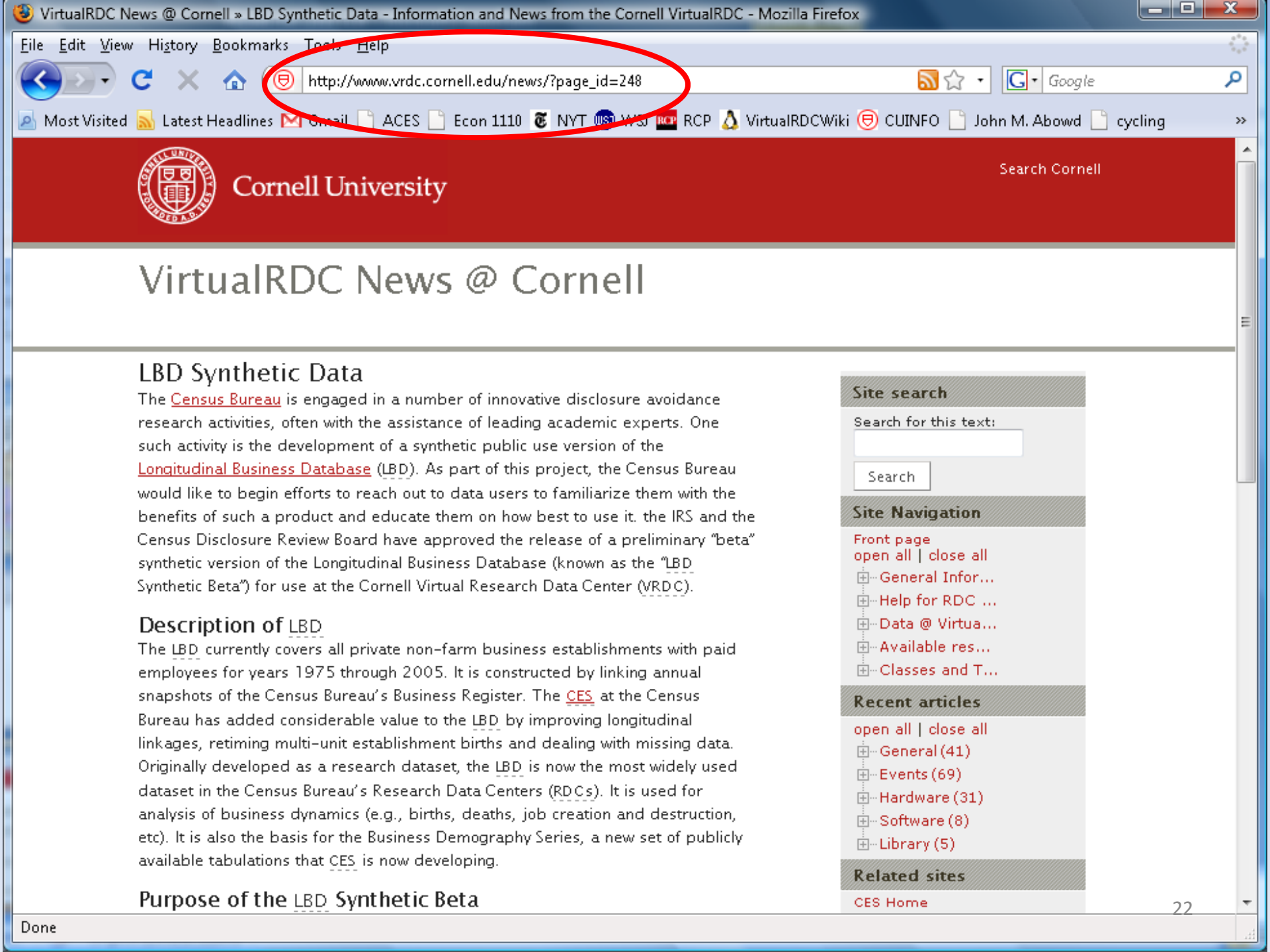
Disclosure Avoidance Review

- Done by probabilistic record linking and distance-based record linking to the source SIPP public use file

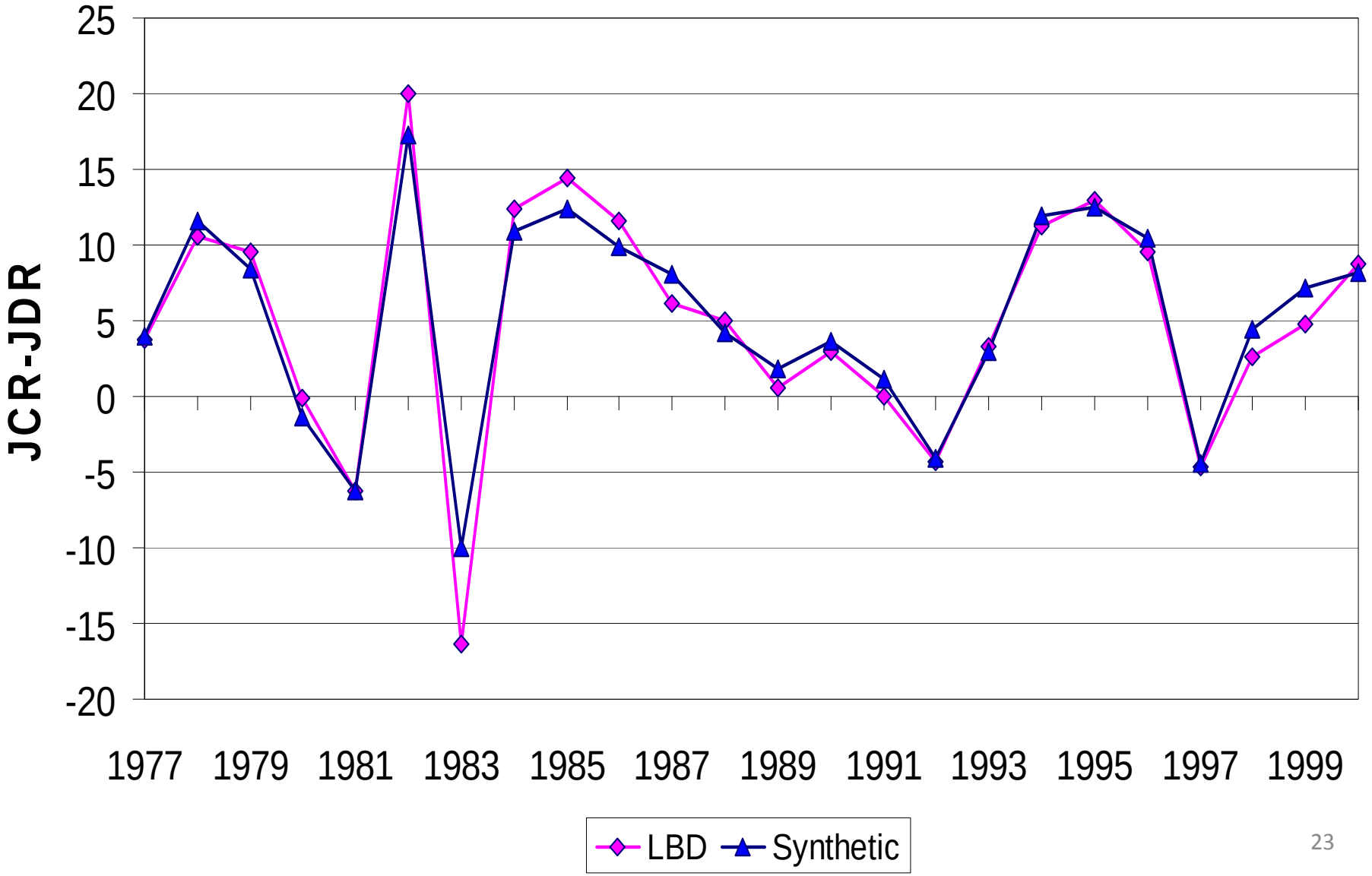
Decile of Distance Distr.	True Matches	False Matches	% True
1	181	26199	0.686%
2	193	26186	0.732%
3	193	26186	0.732%
4	184	26196	0.697%
5	207	26172	0.785%
6	231	26148	0.876%
7	243	26137	0.921%
8	267	26112	1.012%
9	368	26011	1.395%
10	612	25767	2.320%
Overall	2679	261114	1.016%

Project 2: Synthetic Longitudinal Business Database

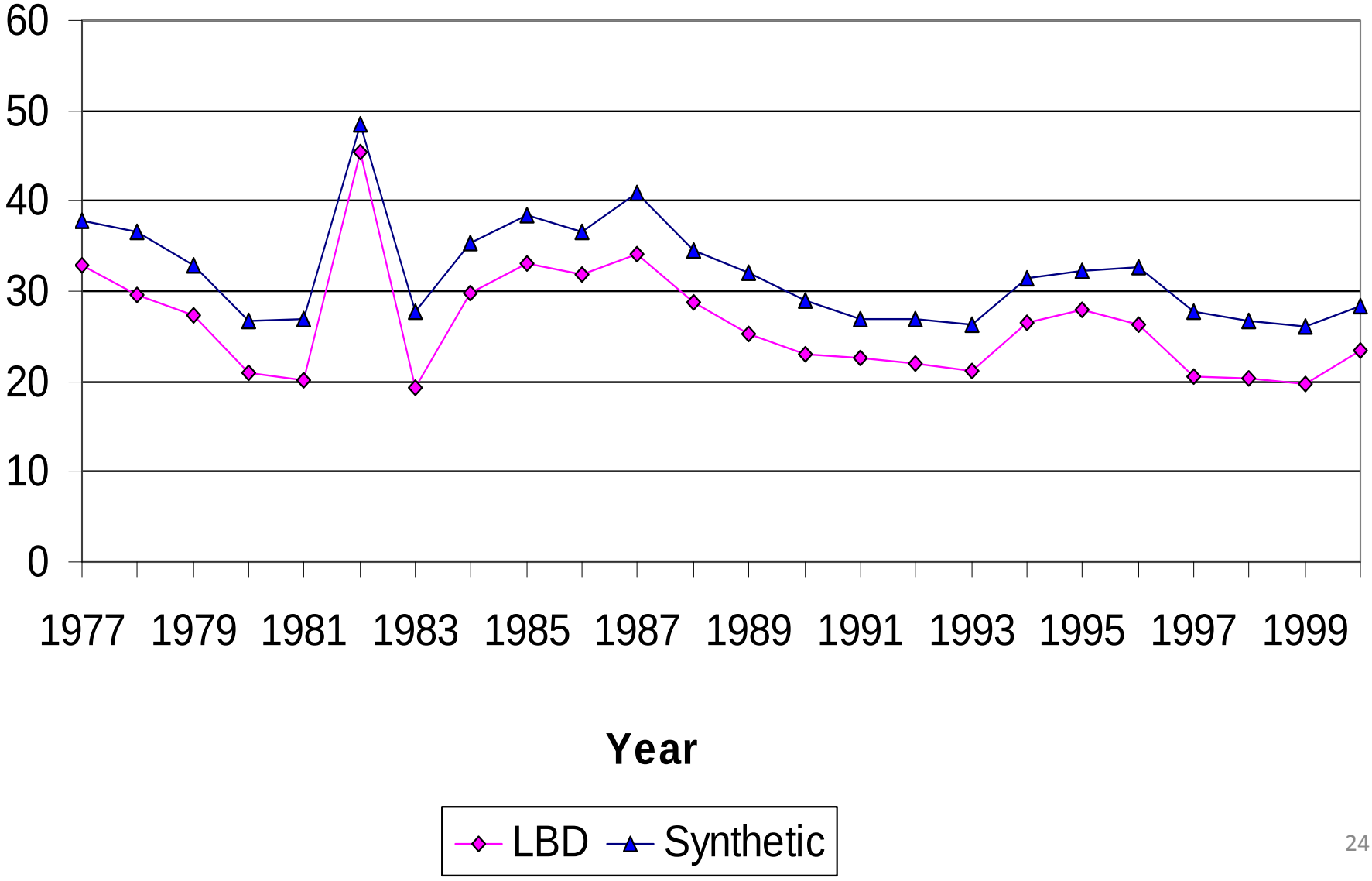
- Census Bureau establishment universe from 1976 to 2001
- Partially synthetic microdata file released in beta for a single industry (SIC 573) in May 2007
- Complete partially synthetic microdata file currently undergoing disclosure avoidance review
- Beta users of the synthetic LBD will be able to have their analyses run on the confidential LBD, as in the SIPP example


VirtualRDC News @ Cornell » LBD Synthetic Data - Information and News from the Cornell VirtualRDC - Mozilla Firefox
File Edit View History Bookmarks Tools Help
http://www.vrdc.cornell.edu/news/?page_id=248
Google
Most Visited Latest Headlines Gmail ACES Econ 1110 NYT WSJ RCP VirtualRDCWiki CUINFO John M. Abowd cycling
Cornell University
Search Cornell
VirtualRDC News @ Cornell
LBD Synthetic Data
The Census Bureau is engaged in a number of innovative disclosure avoidance research activities, often with the assistance of leading academic experts. One such activity is the development of a synthetic public use version of the Longitudinal Business Database (LBD). As part of this project, the Census Bureau would like to begin efforts to reach out to data users to familiarize them with the benefits of such a product and educate them on how best to use it. the IRS and the Census Disclosure Review Board have approved the release of a preliminary "beta" synthetic version of the Longitudinal Business Database (known as the "LBD Synthetic Beta") for use at the Cornell Virtual Research Data Center (VRDC).
Description of LBD
The LBD currently covers all private non-farm business establishments with paid employees for years 1975 through 2005. It is constructed by linking annual snapshots of the Census Bureau's Business Register. The CES at the Census Bureau has added considerable value to the LBD by improving longitudinal linkages, retiming multi-unit establishment births and dealing with missing data. Originally developed as a research dataset, the LBD is now the most widely used dataset in the Census Bureau's Research Data Centers (RDCs). It is used for analysis of business dynamics (e.g., births, deaths, job creation and destruction, etc). It is also the basis for the Business Demography Series, a new set of publicly available tabulations that CES is now developing.
Purpose of the LBD Synthetic Beta
Site search
Search for this text:
Search
Site Navigation
Front page
open all | close all
General Infor...
Help for RDC ...
Data @ Virtua...
Available res...
Classes and T...
Recent articles
open all | close all
General (41)
Events (69)
Hardware (31)
Software (8)
Library (5)
Related sites
CES Home
Done

Net Job Flow Rate



Job Creation Rates: LBD and Implicates by Year



Pearson Correlation Coefficients

SIC 573

Year: 2000

	Employment	Synthetic Employment	Payroll	Synthetic Payroll
Employment	1 41000			
Synthetic Employment	0.003 21100	1 41000		
Payroll	0.712 41000	-0.012 21100	1 41000	
Synthetic Payroll	0.007 21100	0.444 41000	0.004 21100	1 41000

Application to Synthetic LBD

- Categorize each of employment and payroll so that
  - 100 represents observation in 0-90th percentiles
  - 010 represents obs in 90-99th percentiles
  - 001 represents obs in 99+ percentiles
- Construct transition matrix so that
  - 100100 represents point in 0-90 %ile of both employment and payroll
  - 100010 represents point in 0-90th %ile of employment and 90-99th %ile of payroll
  - Etc.

	Synthetic								
Actual	**100100**	**100010**	**100001**	**010100**	**010010**	**010001**	**001100**	**001010**	**001001**
100100	0.8000	0.0250	0.0250	0.0250	0.0250	0.0250	0.0250	0.0250	0.0250
100010	0.0250	0.8000	0.0250	0.0250	0.0250	0.0250	0.0250	0.0250	0.0250
100001	0.0250	0.0250	0.8000	0.0250	0.0250	0.0250	0.0250	0.0250	0.0250
010100	0.0250	0.0250	0.0250	0.8000	0.0250	0.0250	0.0250	0.0250	0.0250
010010	0.0250	0.0250	0.0250	0.0250	0.8000	0.0250	0.0250	0.0250	0.0250
010001	0.0250	0.0250	0.0250	0.0250	0.0250	0.8000	0.0250	0.0250	0.0250
001100	0.0250	0.0250	0.0250	0.0250	0.0250	0.0250	0.8000	0.0250	0.0250
001010	0.0250	0.0250	0.0250	0.0250	0.0250	0.0250	0.0250	0.8000	0.0250
001001	0.0250	0.0250	0.0250	0.0250	0.0250	0.0250	0.0250	0.0250	0.8000

Application Synthetic LBD

- When one obs on one variable changes, it could affect the frequencies in any of the cells of the synthetic data.
- For each of the cells in the synthetic data (columns), calculate the relative probability that two observed data sets could have generated slightly different synthetic data sets.
- Compare only all possible “neighboring” cells in the observed data that could be the source of the difference. These cells differ only by employment or payroll (not both)

Example (Artificial Data)

- Transition matrix (artificial data)
  - Rows represent actual (confidential) data, columns represent synthetic (released) data
  - Rows are the conditional probability of releasing the column value, given the row value
  - The example has been simplified to illustrate the method

	Synthetic								
Actual	**100100**	**100010**	**100001**	**010100**	**010010**	**010001**	**001100**	**001010**	**001001**
100100	0.8000	0.0250	0.0250	0.0250	0.0250	0.0250	0.0250	0.0250	0.0250
100010	0.0250	0.8000	0.0250	0.0250	0.0250	0.0250	0.0250	0.0250	0.0250
100001	0.0250	0.0250	0.8000	0.0250	0.0250	0.0250	0.0250	0.0250	0.0250
010100	0.0250	0.0250	0.0250	0.8000	0.0250	0.0250	0.0250	0.0250	0.0250
010010	0.0250	0.0250	0.0250	0.0250	0.8000	0.0250	0.0250	0.0250	0.0250
010001	0.0250	0.0250	0.0250	0.0250	0.0250	0.8000	0.0250	0.0250	0.0250
001100	0.0250	0.0250	0.0250	0.0250	0.0250	0.0250	0.8000	0.0250	0.0250
001010	0.0250	0.0250	0.0250	0.0250	0.0250	0.0250	0.0250	0.8000	0.0250
001001	0.0250	0.0250	0.0250	0.0250	0.0250	0.0250	0.0250	0.0250	0.8000

Example, Part II

- Differential Privacy Calculation
  - Consider the column 100010
  - Entry highlighted in yellow is the maximum numerator (row value 100010)
  - Entries highlighted in green have row values that differ from the row value for only a single variable (employment or payroll, not both)

	Synthetic								
Actual	**100100**	**100010**	**100001**	**010100**	**010010**	**010001**	**001100**	**001010**	**001001**
100100	0.8000	0.0250	0.0250	0.0250	0.0250	0.0250	0.0250	0.0250	0.0250
100010	0.0250	0.8000	0.0250	0.0250	0.0250	0.0250	0.0250	0.0250	0.0250
100001	0.0250	0.0250	0.8000	0.0250	0.0250	0.0250	0.0250	0.0250	0.0250
010100	0.0250	0.0250	0.0250	0.8000	0.0250	0.0250	0.0250	0.0250	0.0250
010010	0.0250	0.0250	0.0250	0.0250	0.8000	0.0250	0.0250	0.0250	0.0250
010001	0.0250	0.0250	0.0250	0.0250	0.0250	0.8000	0.0250	0.0250	0.0250
001100	0.0250	0.0250	0.0250	0.0250	0.0250	0.0250	0.8000	0.0250	0.0250
001010	0.0250	0.0250	0.0250	0.0250	0.0250	0.0250	0.0250	0.8000	0.0250
001001	0.0250	0.0250	0.0250	0.0250	0.0250	0.0250	0.0250	0.0250	0.8000

Example, Part III

- Differential Privacy Calculation
  - Repeat the calculation for every column of the transition matrix
  - In each column, the yellow entry is the maximum numerator and the green entries are the relevant rows for comparison
  - Differential privacy for this synthesizer is 3.4657
  - The inferential disclosure odds ratio has been bounded at 32

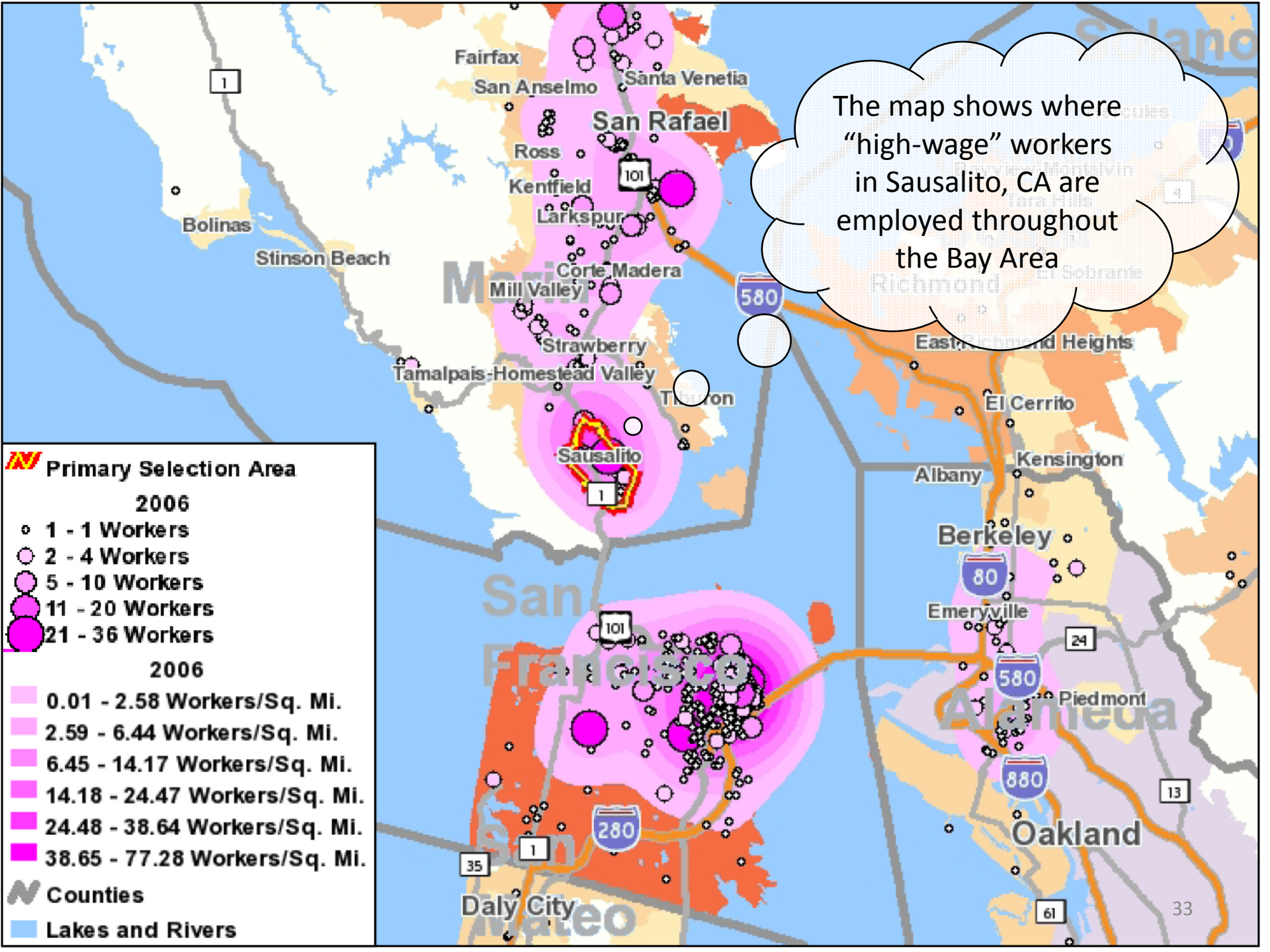
	Synthetic								
Actual	**100100**	**100010**	**100001**	**010100**	**010010**	**010001**	**001100**	**001010**	**001001**
100100	0.8000	0.0250	0.0250	0.0250	0.0250	0.0250	0.0250	0.0250	0.0250
100010	0.0250	0.8000	0.0250	0.0250	0.0250	0.0250	0.0250	0.0250	0.0250
100001	0.0250	0.0250	0.8000	0.0250	0.0250	0.0250	0.0250	0.0250	0.0250
010100	0.0250	0.0250	0.0250	0.8000	0.0250	0.0250	0.0250	0.0250	0.0250
010010	0.0250	0.0250	0.0250	0.0250	0.8000	0.0250	0.0250	0.0250	0.0250
010001	0.0250	0.0250	0.0250	0.0250	0.0250	0.8000	0.0250	0.0250	0.0250
001100	0.0250	0.0250	0.0250	0.0250	0.0250	0.0250	0.8000	0.0250	0.0250
001010	0.0250	0.0250	0.0250	0.0250	0.0250	0.0250	0.0250	0.8000	0.0250
001001	0.0250	0.0250	0.0250	0.0250	0.0250	0.0250	0.0250	0.0250	0.8000

Project 3: Synthetic Longitudinal Employer-Household Dynamics Data

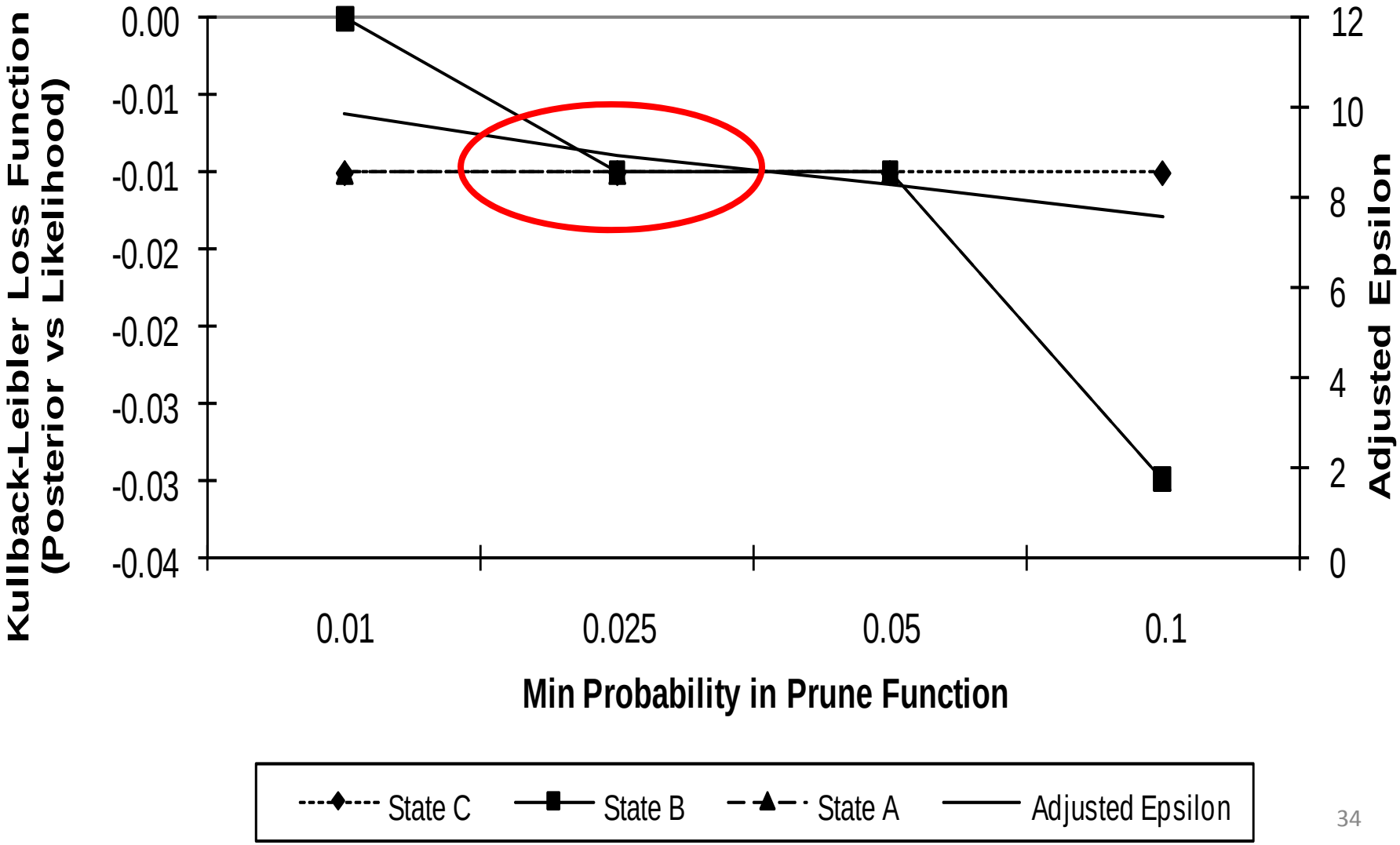
- You saw the presentation yesterday by Simon
- Longitudinally integrated relational database of individuals, employers, jobs
- Complete realized mobility graph (who ever worked for whom) not synthesized.

Related Projects: OnTheMap

- Residential location-workplace location mapping application developed from the Census Bureau's Longitudinal Employer-Household Dynamics infrastructure file system
- Block-level data for 2002 through 2006
- Complete origin-destination matrix stratified by age group, income group, and industry protected by synthetic data methods


The map shows where "high-wage" workers in Sausalito, CA are employed throughout the Bay Area
Primary Selection Area
2006
1 - 1 Workers
2 - 4 Workers
5 - 10 Workers
11 - 20 Workers
21 - 36 Workers
2006
0.01 - 2.58 Workers/Sq. Mi.
2.59 - 6.44 Workers/Sq. Mi.
6.45 - 14.17 Workers/Sq. Mi.
14.18 - 24.47 Workers/Sq. Mi.
24.48 - 38.64 Workers/Sq. Mi.
38.65 - 77.28 Workers/Sq. Mi.
Counties
Lakes and Rivers
Fairfax
San Anselmo
Santa Venetia
San Rafael
Ross
Kentfield
Larkspur
Bolinas
Stinson Beach
Corte Madera
Mill Valley
Strawberry
Tamalpais-Homestead Valley
Sausalito
El Cerrito
Albany
Kensington
Berkeley
Emeryville
Piedmont
Oakland
Daly City
1
101
580
80
280
880
24
13
35
61

Analytical Validity v. Confidentiality Protection in on the Map



Taking Account of Formal Privacy Models

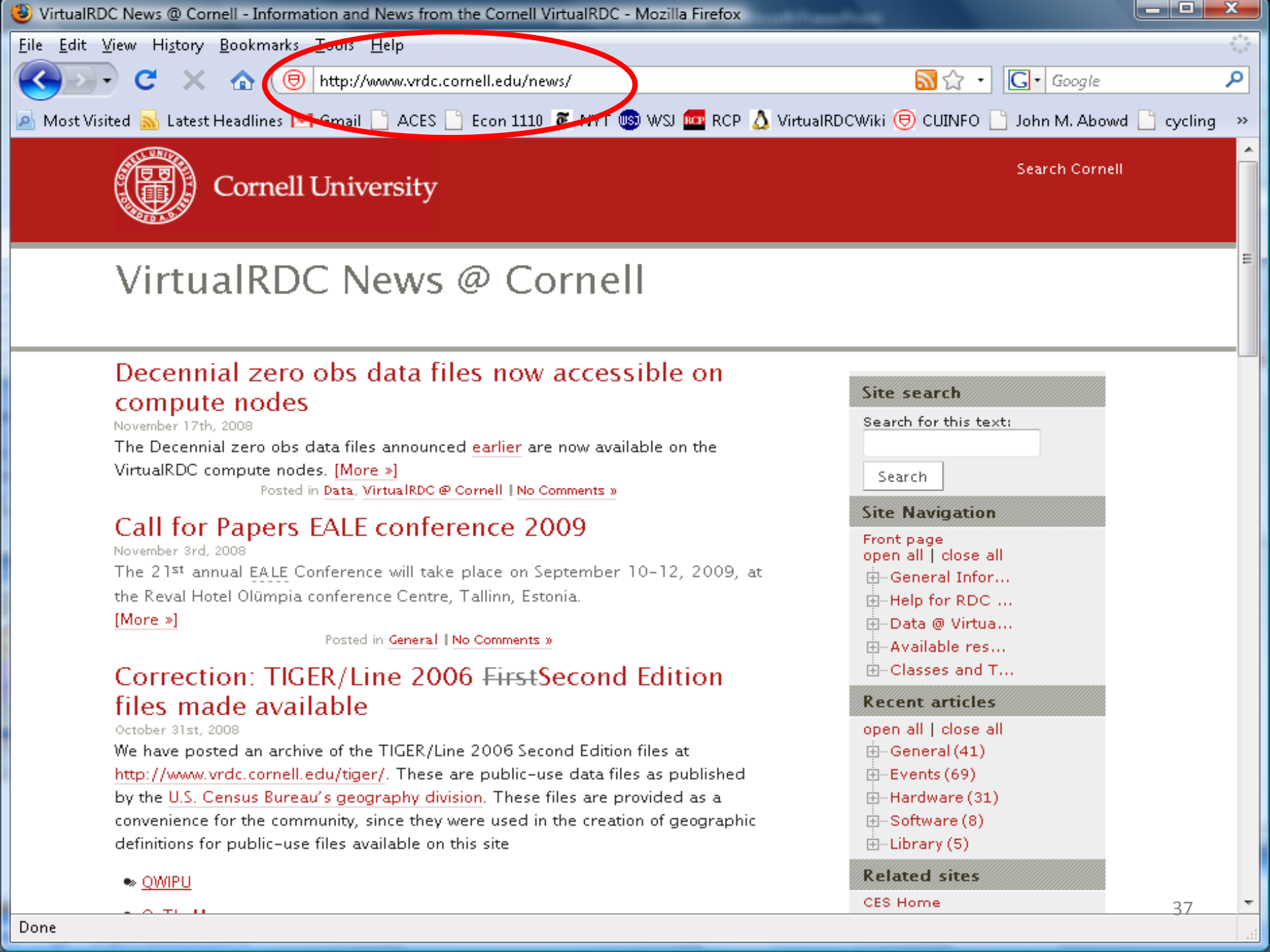
- It is now known from a variety of papers in the cryptographic data privacy literature (Dwork, Nissim and their many collaborators, Gehrke and his collaborators, and others) that the confidentiality protection afforded by synthetic data depends upon properties of the transition probabilities that relate the confidential data to the release data.
- Not surprising since

$$\Pr[\tilde{X}|X] = I$$

implies that the PPD leaves the confidential data unchanged by the synthesizer.

The Virtual Research Data Center

- Web-based application that allows complete access to all synthetic data products developed by the NSF-ITR collaborators with the Census Bureau
- Full access from anywhere
- Simulates the Census RDC operation to facilitate developing models on synthetic data that can be estimated on the underlying confidential data


VirtualRDC News @ Cornell - Information and News from the Cornell VirtualRDC - Mozilla Firefox
File Edit View History Bookmarks Tools Help
http://www.vrdc.cornell.edu/news/
Google
Most Visited Latest Headlines Gmail ACES Econ 1110 NYT WSJ RCP VirtualRDCWiki CUINFO John M. Abowd cycling
Cornell University
Search Cornell
VirtualRDC News @ Cornell
Decennial zero obs data files now accessible on compute nodes
November 17th, 2008
The Decennial zero obs data files announced earlier are now available on the VirtualRDC compute nodes. [More »]
Posted in Data, VirtualRDC @ Cornell | No Comments »
Call for Papers EALE conference 2009
November 3rd, 2008
The 21st annual EALE Conference will take place on September 10-12, 2009, at the Reval Hotel Olümpia conference Centre, Tallinn, Estonia.
[More »]
Posted in General | No Comments »
Correction: TIGER/Line 2006 FirstSecond Edition files made available
October 31st, 2008
We have posted an archive of the TIGER/Line 2006 Second Edition files at http://www.vrdc.cornell.edu/tiger/. These are public-use data files as published by the U.S. Census Bureau's geography division. These files are provided as a convenience for the community, since they were used in the creation of geographic definitions for public-use files available on this site
QWIPU
Site search
Search for this text:
Search
Site Navigation
Front page
open all | close all
General Infor...
Help for RDC ...
Data @ Virtua...
Available res...
Classes and T...
Recent articles
open all | close all
General (41)
Events (69)
Hardware (31)
Software (8)
Library (5)
Related sites
CES Home
Done

Contacts

- John.Abowd@cornell.edu
- virtualrdc@cornell.edu

Thank you