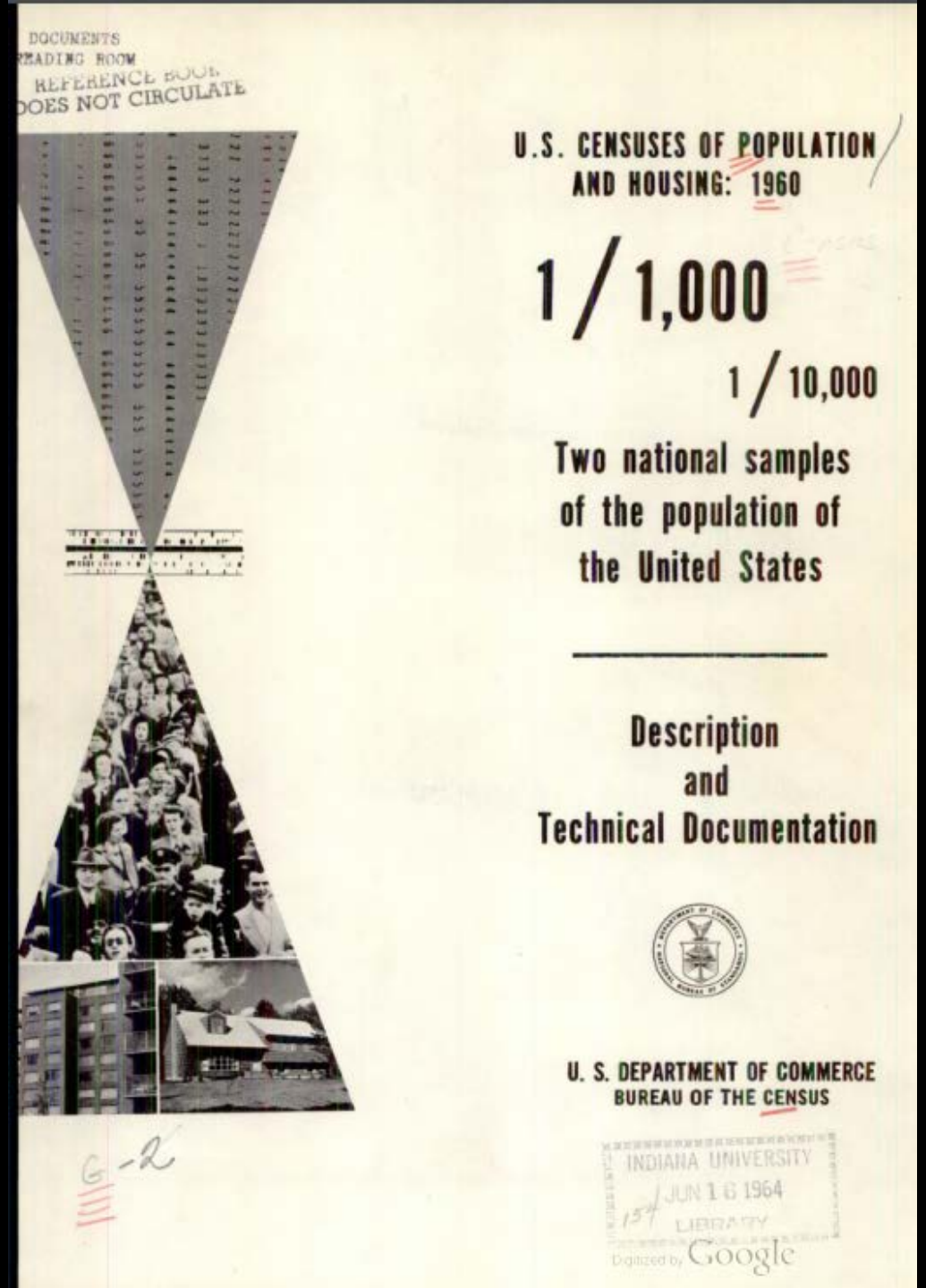
Challenges of Synthetic ACS microdata

Steven Ruggles
University of Minnesota



1960
The invention of microdata

To preserve confidentiality, the Census Bureau removed names, details on place of residence, and other potentially identifying information. The Bureau explained that “therefore, it has been determined that making records available in this form does not violate the provision for confidentiality in the law under which the census was conducted”

Otis Dudley Duncan

“The importance of this innovation can hardly be overestimated... With access to the unit records, the social scientist may specify in detail how variables are to be manipulated so as to produce an optimal estimate of the magnitude desired”

Power of Public Use Microdata

- The Census Bureau has released Public Use Microdata Samples for long form data from each decennial census since 1960, and for the ACS since 2000 after it replaced the long form census

- The 1960-2024 microdata include information on 140 million persons nested into 51 million households

- ACS microdata are the single most intensively-used data source in social science and policy research:
  - Over 50,000 Google Scholar citations
  - Over 5,400 dissertations
  - IPUMS has fielded 2.7 million requests from 250,000 investigators, currently 21,000 data requests per month

Power of Public Use Microdata

- The ACS microdata are core research infrastructure for analysis of poverty, inequality, immigration, internal migration, race, ethnicity, disability, health insurance, transportation, housing, fertility, nuptiality, marital instability, work, education, and family composition.

- They are analogous to a Hubble Telescope for social science, although the ACS is a lot cheaper than Hubble and has yielded a far larger number of scientific publications

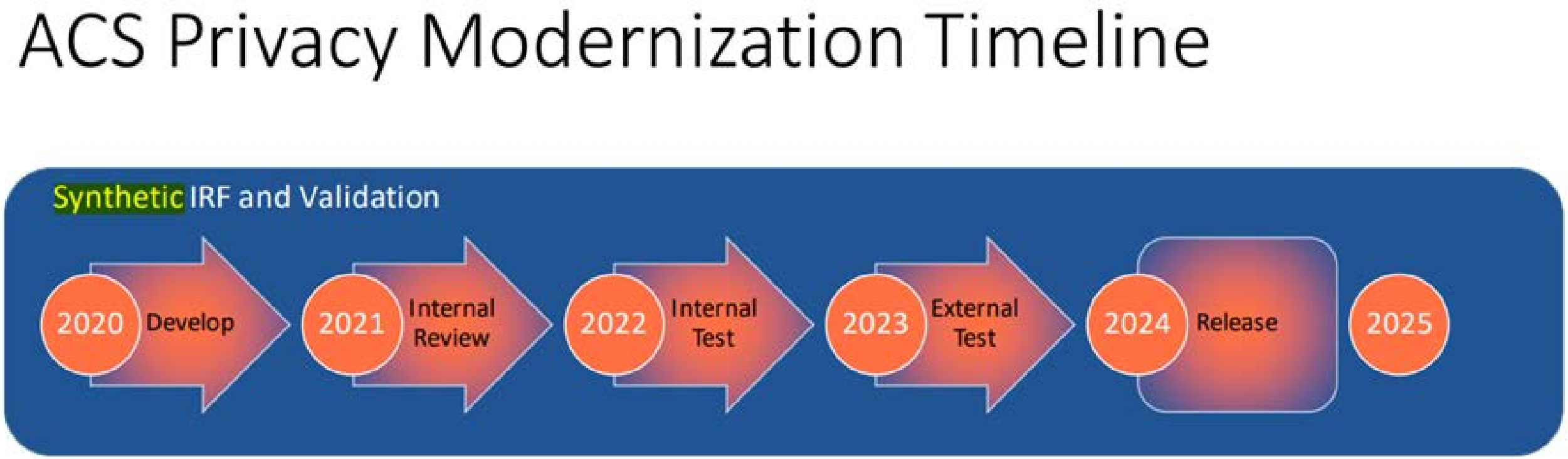
Existing disclosure control for ACS microdata

A. Methods that preserve data integrity (Skinner et al. 1994)
    Sampling
    Suppression of detail

B. Contamination methods
    Swapping
    Age perturbation (large units and age 65+)
    Partially synthetic group quarters

ACS Data Users Conference, May 20, 2021:

Rolando Rodríguez described Census Bureau plans to replace the American Community Survey (ACS) microdata with “fully synthetic” data by 2024.



Abowd et al. (2020) describes the approach as follows:

1. Build a chain of models, simulating each variable successively given the previous synthesized variables (Raghunathan et al., 2001). Currently, the team is assessing the use of classification trees for this purpose (Reiter, 2005);
2. Create synthetic microdata from these models for all records and all variables, creating fully synthetic data; and
3. Allow users to validate results from the synthetic microdata against the internal data. Validated results would have to meet the same standards for disclosure avoidance as all other public data releases and would be limited in quantity to statistics required for the stated purpose.

The ACS has a hierarchical structure, with persons nested within households.

The interrelationships of variables across household members are extremely complex, and the Census Bureau has struggled to capture them in their models.

Formal Privacy and Synthetic Data for the American Community Survey

Michael H. Freiman

U.S. Census Bureau[1]

Rolando A. Rodríguez

U.S Census Bureau[1]

Jerome P. Reiter

Duke University and U.S. Census Bureau[1]

Amy Lauger

U.S. Census Bureau[1]

Abstract: The U.S. Census Bureau is expanding the use of formal privacy to improve disclosure avoidance methods and to enable quantification of privacy loss. This paper discusses the particular challenges of applying formally private algorithms to the American Community Survey (ACS), including the data's high dimensionality coupled with sample size limitations and the use of complex survey weights. We describe research on model-based approaches to creating synthetic data for the ACS, focusing on health insurance.

Formal Privacy

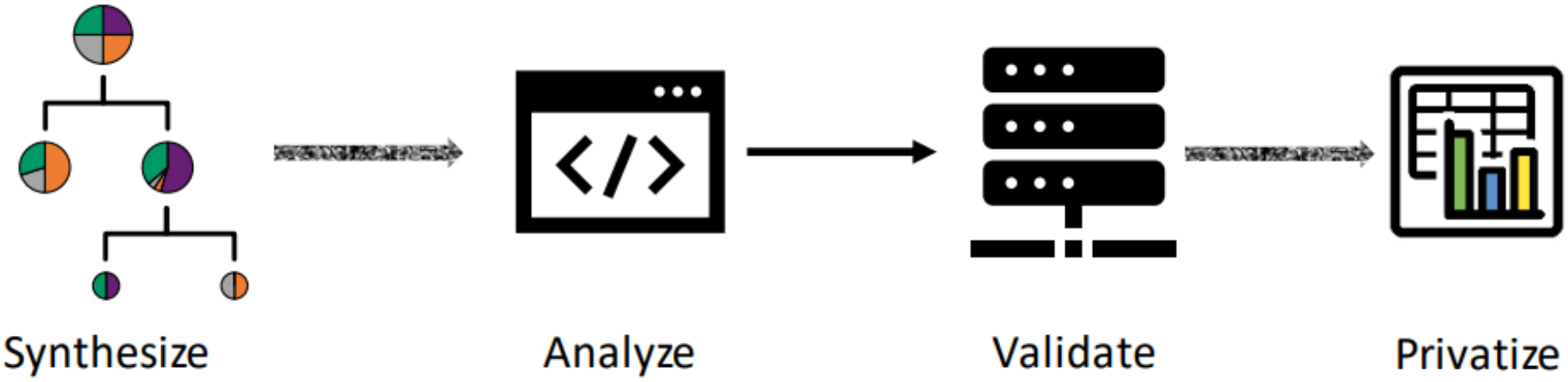
For several decades, the Census Bureau and other data providers have used disclosure avoidance methods to protect the data provided by individual respondents. Such methods are necessary because merely removing direct identifiers—what would traditionally be considered personally identifiable information—is insufficient to prevent the identity or attributes of a respondent from being inferred, either with certainty or with a high degree of confidence. The methods used have varied by data product, but the main traditional method for household records in the American Community Survey (ACS) has been data swapping, wherein pairs of records that are identical with

Conclusion:

This research indicates that the synthesis methods we are currently using do not preserve (or approximately preserve) all of the multivariate relationships that we might like to see preserved, in particular relationships between insurance status and relationship to householder or between different types of insurance status.

The Census Bureau acknowledges that the synthetic microdata will be unsuitable for research purposes.
Users will instead validate their results by submitting their code to the Census Bureau, which would run the code on real data, conduct disclosure review on the output, and provide “true” results to the user.

- Making synthetic data to satisfy all use cases is impossible
- Users will be able to validate synthetic output against internal data



The best-known previous example of synthetic population microdata released by the census Bureau: The SIPP Synthetic Beta File, which combines data from the Survey of Income and Program Participation with data from Social Security.

The data were disseminated by the VirtualRDC at Cornell.

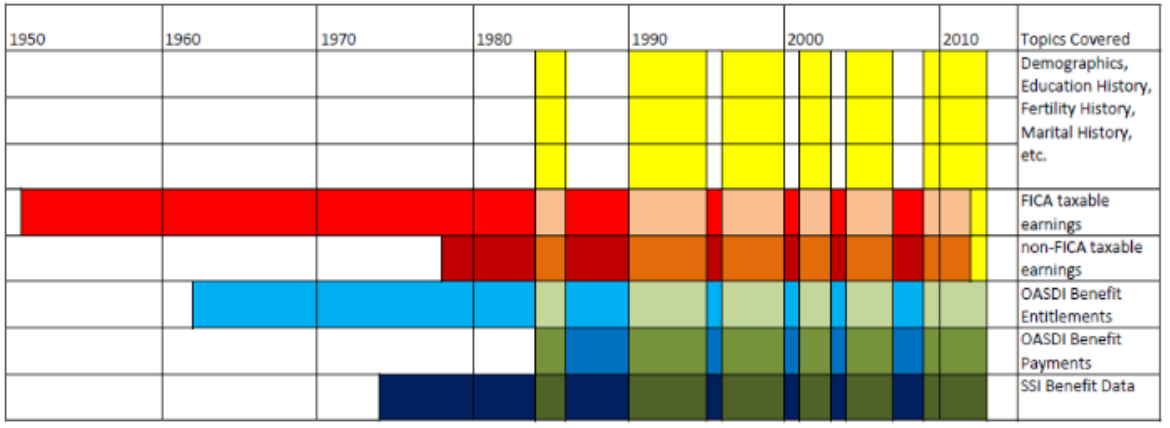
SIPP Synthetic Beta file

On this page

Print

Description

The SIPP Synthetic Beta (SSB) is a **Census Bureau** product that integrates person-level micro-data from a household survey with administrative tax and benefit data. These data link respondents from the **Survey of Income and Program Participation** (SIPP) to Social Security Administration (SSA)/Internal Revenue Service (IRS) Form W-2 records and SSA records of retirement and disability benefit receipt, and were produced by Census Bureau staff economists and statisticians in collaboration with researchers at Cornell University, the SSA and the IRS. The following graph describes succinctly the availability and coverage of survey responses and administrative data:



Synthetic SIPP usage

- About 170 users over the first seven years the data were available
- Approximately 15 of these users had results validated (about 2 per year)
- Only a handful of substantive research publications have appeared

The ACS generates at least 100,000 times as much traffic as the synthetic SIPP.

If all those users turned to a synthetic ACS, the Census Bureau would have to carry out tens of thousands of validations each year, and tens of thousands of disclosure evaluations of the results.

End of life for the Cornell Synthetic Data Server September 30, 2022

(posted by Lars Vilhuber, July 20th, 2022)

The Cornell Synthetic Data Server, in its current format, has been used by over 300 researchers since we first started it in 2014 (8 years ago!). It has, I believe, been a very useful tool to get researcher feedback on the quality of the data, and has allowed researchers to do terrific work.

Print

Unfortunately, the last external funding to support the SDS ended several years ago, and the server has been supported through funds from a now-emeritus Cornell faculty member. The servers are old, and no longer have active maintenance.

> It is therefore with a heavy heart that I am announcing the shutdown of the current servers on **September 30, 2022**.

The respective data providers at the Census Bureau are looking into alternative options, including replacement server access and novel validation methods.

Please continue to submit project proposals, to identify the demand for this type of service. However, new accounts on the Cornell Synthetic Data Server will no longer be routinely created.

We thank all of our funders over the years, including NSF grants **SES-1042181** and **BCS-0941226**, a grant from the **Alfred P. Sloan Foundation**, and the **Edmund Ezra Day chair** at the Cornell ILR School.

Lars Vilhuber, Executive Director, LDI
*John M. Abowd, **Edmund Ezra Day** Professor Emeritus of **Economics**, Statistics and Data Science.*

Synthetic SIPP Data

ANNOUNCEMENT

Update 07/08/2024: We now have a new process in place! Please read below the description and instructions for the process to access the SSB data and validation requests. Please note that it might take longer than usual to address requests initially, as we work on setting up both previous and new SSB users in the new system.

The reason the Census Bureau wants to eliminate one of the world's most intensively used scientific resources is concern about respondent confidentiality.

The Rodriguez presentation acknowledged that there has never been a documented case of identification of a respondent in the ACS census microdata, but he argued that there are unknown risks.

No public disclosures from Census products does not equate to no privacy concerns… or risks

2012 Tenth Annual International Conference on Privacy, Security and Trust

Exploring re-identification risks in public domains

Aditi Ramachandran
Georgetown University
ar372@georgetown.edu

Lisa Singh
Georgetown University
singh@cs.georgetown.edu

Edward Porter
US Census Bureau
edward.h.porter@census.gov

Frank Nagle
Harvard Business School
fnagle@hbs.edu

***Abstract*—While re-identification of sensitive data has been studied extensively, with the emergence of online social networks and the popularity of digital communications, the ability to use public data for re-identification has increased. This work begins by presenting two different cases studies for sensitive data re-identification. We conclude that targeted re-identification using traditional variables is not only possible, but fairly straightforward given the large amount of public data available. However, our first case study also indicates that large-scale re-identification is less likely. We then consider methods for agencies such as the Census Bureau to identify variables that cause individuals to be vulnerable without testing all combinations of variables. We show the effectiveness of different strategies on a Census Bureau data set and on a synthetic data set.**

I. Introduction

With the emergence of online social networks and social media sites, the increase in Web 2.0 content, and the popularity of digital communication, more and more public information about individuals is available on the Internet. While much of this information is not sensitive, it is not uncommon for users to publish some sensitive information, including birth dates and addresses on social networking sites. When potential adversaries have access to this large corpus of publicly available and potentially sensitive data, abuse can take place, leading to consequences such as fraud, stalking, and identity theft.

Cynics may question the value of protecting sensitive information that users are readily making public. Even so, a need to protect potentially sensitive data still exists for government entities and corporations. When agencies, such as the Census Bureau, release survey data, they need to be confident that the data cannot be used to re-identify survey participants. Not only are some data fields sensitive, e.g. income, but fewer individuals will participate in surveys truthfully if they are not confident that their identities will be protected.

This paper begins by presenting different strategies for re-identification. To better understand the level of difficulty associated with linking public information to other public data or to anonymized public information, we conducted two case studies, one involving Census Bureau data and one involving social networking data. The goals of each case study are literature, we conclude that targeted re-identification is not only possible, but relatively straightforward given the large amounts of publicly available data.

It would be beneficial for agencies, such as the Census Bureau to have a suite of tools to help them find these vulnerable individuals and distinct combinations of attribute features. Unfortunately, for large data sets containing a large number of attributes and a large number of records, exhaustively searching for individuals that are targets for re-identification is very costly. We present different heuristics that attempt to accurately identify attribute feature combinations causing individuals to be vulnerable without exhaustively searching all combinations and show their effectiveness on a Census Bureau data set and on synthetic data.

The contributions of this paper are as follows: (1) we present two different re-identification strategies using real world data sets; (2) we compare strategies for identifying variables that are causing individuals to be reidentified in synthetic data; (3) we analyze these strategies on real world data and conclude that the effectiveness of the strategies is very dependent on the type of data vulnerability present.

The remainder of this paper is organized as follows. Section II presents relevant literature. Section III presents a re-identification case study using public Census Bureau data, while section IV presents one using Twitter and Facebook data. In section V, strategies for finding variables that cause the vulnerability without testing all variable combinations are explored. Conclusions are presented in section VI.

II. Related Literature

A current area of research that often applies to crimes such as identity theft and fraud involves re-identification, or the process by which anonymized personal data is matched with its true owner. Even though potentially sensitive data is typically anonymized, re-identification approaches can sometimes be used to discover the identity of certain people in a data set [2], [1], [13]. Sweeney used a purchased voter registration list for Cambridge, Massachusetts and a publicly available

2012 Tenth Annual International Conference on Privacy, Security and Trust

Exploring re-identification risks in public domains

Aditi Ramachandran
Georgetown University
ar372@georgetown.edu

Lisa Singh
Georgetown University
singh@cs.georgetown.edu

Edward Porter
US Census Bureau
edward.h.porter@census.gov

Frank Nagle
Harvard Business School
fnagle@hbs.edu

***Abstract*—While re-identification of sensitive data has been studied extensively, with the emergence of online social networks and the popularity of digital communications, the ability to use public data for re-identification has increased. This work begins by presenting two different cases studies for sensitive data re-identification. We conclude that targeted re-identification using traditional variables is not only possible, but fairly straightforward given the large amount of public data available. However,**

Concluded that 0.017 percent of respondents were vulnerable to possible re-identification, but 78 percent of those putative re-identifications were false, and an outsider would have no means of determining which ones were correct.

versaries have access to this large corpus of publicly available and potentially sensitive data, abuse can take place, leading to consequences such as fraud, stalking, and identity theft.

Cynics may question the value of protecting sensitive information that users are readily making public. Even so, a need to protect potentially sensitive data still exists for government entities and corporations. When agencies, such as the Census Bureau, release survey data, they need to be confident that the data cannot be used to re-identify survey participants. Not only are some data fields sensitive, e.g. income, but fewer individuals will participate in surveys truthfully if they are not confident that their identities will be protected.

This paper begins by presenting different strategies for re-identification. To better understand the level of difficulty associated with linking public information to other public data or to anonymized public information, we conducted two case studies, one involving Census Bureau data and one involving social networking data. The goals of each case study are

literature, we conclude that targeted re-identification is not only possible, but relatively straightforward given the large amounts of publicly available data.

It would be beneficial for agencies, such as the Census Bureau to have a suite of tools to help them find these vulnerable individuals and distinct combinations of attribute features.

very dependent on the type of data vulnerability present.

The remainder of this paper is organized as follows. Section II presents relevant literature. Section III presents a re-identification case study using public Census Bureau data, while section IV presents one using Twitter and Facebook data. In section V, strategies for finding variables that cause the vulnerability without testing all variable combinations are explored. Conclusions are presented in section VI.

II. Related Literature

A current area of research that often applies to crimes such as identity theft and fraud involves re-identification, or the process by which anonymized personal data is matched with its true owner. Even though potentially sensitive data is typically anonymized, re-identification approaches can sometimes be used to discover the identity of certain people in a data set [2], [1], [13]. Sweeney used a purchased voter registration list for Cambridge, Massachusetts and a publicly available

More Research is Needed

1. New Reidentification studies

The Census Bureau should conduct a new generation of re-identification to identify realistic vulnerabilities that have the potential to allow positive identification of respondents.

If risks are identified, we need targeted solutions to address them. The one study we do have suggests that risk is confined to a small group of outliers and could be addressed with minimal harm to data integrity by focusing on identifying variables in this subgroup.

More Research is Needed

2. Studies of the scientific impact of new disclosure controls

Any changes to disclosure control should be thoroughly assessed before those changes are implemented.

The best way to assess usability is to replicate a wide range of past published studies using versions of the data that incorporate disclosure controls under consideration.

The user community can do this work. For this to happen, the Census Bureau would need to prepare test files that apply new techniques to older ACS survey years, so investigators can replicate past analyses and assess usability of the data.

New research can optimize tradeoffs between usability and risk

The combination of new research on disclosure risk and new research on the impact of disclosure controls can inform the development of effective methods that optimize the tradeoffs between utility and risk.

Partially synthetic methods that focus on identifying variables (those available in external data sources) and vulnerable records (with rare combinations of identifying variables) offer a promising avenue to enable continued access to high-quality public use microdata.

The benefits of eliminating public access to usable ACS microdata are highly uncertain.

The costs, however, are clear:
Eliminating access to ACS microdata would severely damage the nation’s statistical infrastructure;
there is no substitute for most applications.

The Power of Public Data

In 1959, Conrad Taeuber proposed the first microdata sample:

“It is essential to this proposal that these materials would be ***equally available to anyone***” and that the user “would be free to use them or subsamples of them as he saw fit.”

Public Data are a Public Good

- It matters whether data are accessible to a broad public or are restricted to a narrow elite.
- If reliable data become inaccessible or available only to select investigators, it will at minimum make evidence-based research more expensive and less reproducible; it will make some research impossible.
- If the government can pick and choose which researchers get access to data, that is a core threat to science.

PNAS PERSPECTIVE

The shortcomings of synthetic census microdata

Steven Ruggles[a,b,1]

Edited by Richard Alba, City University of New York, Graduate Center, New York, NY; received December 31, 2024; accepted January 21, 2025

U.S. Census Bureau officials recently reaffirmed the Bureau's ongoing efforts to replace the American Community Survey (ACS) public use microdata sample with "fully synthetic" data to protect respondent confidentiality. With the growth of computing power and expansion of private sector data about the population, the Census Bureau has valid concerns about confidentiality threats to public data. The current plan for fully synthetic data, however, threatens a cornerstone of the nation's scientific infrastructure. Census microdata samples are among the most frequently used sources in social science research, and they are an essential tool for policy formation and planning from the local to the national level. Synthetic census microdata are not suitable for most research and policy applications. There have been no recent attempts to quantify disclosure risk in the ACS microdata, and the sole existing study failed to establish a credible

for the purpose at hand. With access to the unit records, the social scientist may specify in detail how variables are to be manipulated so as to produce an optimal estimate of the magnitude desired" (4).

For the 1970 Census, the Bureau dramatically increased the sample size and improved the level of detail provided in the microdata, especially for geographic information. In 1973, the Bureau also developed an expanded version of the 1960 microdata sample designed to be as compatible as possible with the 1970 census public use samples. The availability of these files unleashed a wave of new research on demographic and economic change in the decade between 1960 and 1970. In response to this success, the Census Bureau produced similar large microdata samples for the censuses of 1980, 1990, and 2000, and they became a mainstay of social and economic research (5).