Stephen H Cohen

NORC at the University of Chicago

Discussion: Formally Private Disclosure Limitation Methods for Establishment Data

Introduction

This session was focused on sorely needed research on release public data sets for establishment data. Current deidentification techniques can result in suppressing over 50% of the estimates. Also, to date establishment surveys have no publicly available datasets.

The research application is the Bureau of Labor Statistics' (BLS) Quarterly Current Employment and Wage Dataset (QCEW). Release of more publicly available data would greatly enhance research possibilities. Research in this session used formal differential privacy techniques that have been successfully applied to Census 2020. The session had three papers: the first paper presented by Daniel Kifer, Penn State University, focused on challenges to introducing formal privacy techniques. The second paper presented by Scott Holan, University of Missouri/U.S. Census Bureau, discussed how to fill in gaps caused by the granularity introduced using formal privacy techniques by developing models based upon universe datasets. The final paper is really the apex of this session. The paper develops formal privacy techniques that are applied to QCEW.

Issues in Releasing Establishment Data

There are no public use files released for establishment as up to now no safe method to release quality data exists while minimizing reidentification risk. Availability of public use file would be a big advance even if they have limitations since today current access is only available at best in:

- restricted to data centers
- remote data access
- data licensees.

Current techniques mainly use cell suppression to minimize reidentification. Th usual rules used in with cell suppression techniques require a minimum number of units in each publishable cell and no single unit that provides a significant amount of data to the cell. See the U.S. Office of Management and Budgets' Working Paper Number 22. The consequences of these rules are significant secondary cell suppression and/or data roll ups.

The research in this session explores whether formal differential privacy techniques (DP) can work to minimize reidentification for establishment data and produce quality data. The main issue with DP techniques is deciding the tradeoff between data protection versus data quality. How would this play out with cells with small counts or dominant units?

Specific issues to be addressed:

- Cells with small counts be publishable.
    - Detail --everybody wants

    - Will establishment data be protected in a PUF
- Can DP protect from reidentification data from units that are easily identifiable such as:
    - Boeing in Seattle
    - Rock quarries in Montgomery County MD which is mostly a white collar
    - Large auto plant in rural Trumbell County Ohio

A key decision in using formal privacy techniques is the tradeoff between the level of privacy protection versus data quality. Data quality results in loss of granular detail which is critical to be displayed for any proposed new methodology.

Quarterly Census of Employment and Wages (QCEW)

QCEW is a Federal State Cooperative Program. QCEW data is collected in partnership with the States and the District of Columbia. Data is derived from unemployment tax returns. Data reported for each physical location by individual county and reported to BLS with monthly employment totals and quarterly wage totals. Data is reported to BLS quarterly. Each physical location is coded to a 6-digit North American Industry Classification System (NAICS) code. BLS releases estimates for each 3143 county in the US, each state and US totals. More than 50% of the data cells are not released due to reidentification concerns and secondary suppressions.

At the most detailed level a cell is a 6 digit NAICS code by an individual county.

Many counties have only 1 or 2 establishments for a unique NAICS code or a single establishment contributing significant data to the estimate.

BLS has gone on record that it can not protect existence of an establishment or its NAICS code.

Summary

The Kifer paper is an excellent discussion of issues associated with releasing fully protected QCEW data that minimizes reidentification. The issues range for the huge amount of data suppressed due to secondary suppression to protect identifiable data to the inconsistencies caused by the Staes using different strategies than those that must be applied by BLS to the national QCEW data.

The Holan paper: "Relating Legacy Methods to Formal Privacy and Leveraging Statistical Modeling in the Release of Formally Private Data" developed Bayesian models to fill in estimated data that cannot be released using formal privacy techniques. The research data set was American Indian Alaska Native (AIAN) American Community Survey (ACS) data in Oklahoma for tribal areas that cross counties or fail quality measures. The models were able to develop estimates for these areas by using Census data.

The paper did not directly do simulations on establishment data but inferred the methods could be extended which I think might be possible but needs to be explored.

These methods would not work if there are no large data sets that be used in the models. It would not work for QCEW as the only equivalent dataset is the Census data frame which is protected by Title 13 of the United States code. Care must be taken in developing appropriate models.

The final paper: "Valid Statistical Methods for Establishment Data Protected with Formally Private Methods" is the key paper in this session. Formal privacy developed in the paper aims to: provide measurable and adjustable privacy-protections, protect against attacks with outside information about the data, and allow more transparency in the process than previous methods such as cell suppression. The proposed mechanism works by adding Gaussian noise followed by post processing to dataset units. The mechanism produces privacy-protected microdata by minimizing a weighted sum of squared differences which uses the noisy values from each establishment and noisy aggregate values for key queries with the variance plug-in estimator as inverse weights. The paper goes through the process using simulated data.

Conclusion

Papers in this session explore reidentification for establishment data using formal differential privacy algorithms. The key paper shows how formal privacy can work with establishment data using a plug in estimator on noisy values of microdata and aggregate estimates.

A second work in this session proposes solutions to granularity problems associated with application of formal differential privacy methodologies.

Two key questions that need to be debated given BLS being on record that it cannot protect the existence or business activity of an establishment:

- Can formal privacy methods yield quality deidentified data with few (e.g. 3) establishments in a county with 6 digit NAICS code that met user needs?
- Can formal privacy add enough noise to unique establishments protect wage and employment estimates adequately while meeting quality parameters without some collapsing?

For Holan's Bayesian models, can we find appropriate data sets to model loss of granularity using formal **privacy mechanisms?**