Integrating Disclosure Limitation into TSE

Alan F. Karr

National Institute of Statistical Sciences
Research Triangle Park, NC 27709 USA

ITSEW 2013, Ames, IA, June 2, 2013

NISS
The Statistics Community Serving the Nation

Why Should Anyone Care?

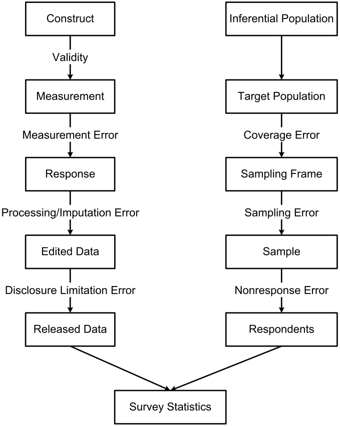
Statistical disclosure limitation (SDL) is the one step in the survey process where error is *introduced deliberately*, for the sake of protecting respondent privacy and dataset confidentiality

To date, there is a *disconnect*

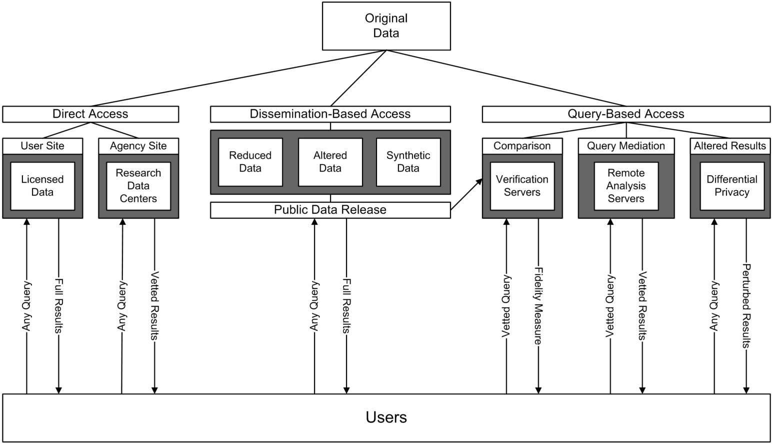
- SDL does not account for other sources of error, especially measurement error
- Efforts to reduce other sources of error do not account for SDL
- Efforts to unify edit, imputation and SDL are few

MY POINT: THIS NEEDS TO CHANGE

Where SDL Fits



High-Level View of SDL



Additive Noise

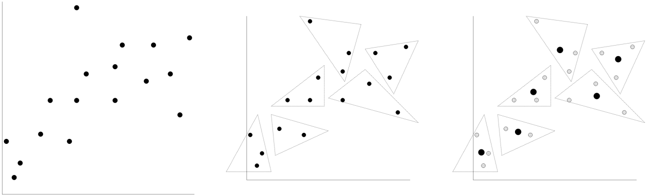
Statistical approach Add noise *to data* prior to release, preserving low-dimensional structure but obscuring high-dimensional, confidentiality-threatening details

Computer science approach, aka differential privacy In a server setting, add noise *to query results*, with verifiable level of protection

Microaggregation

Assume k-dimensional numerical data

1. Group data into sets of size m ($m = 3$ is typical)
2. Replace elements of each m-tuplet by attribute-wise mean



Original Data The Clusters The End Product

Combined Methods

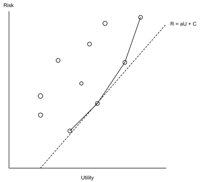
No reason to use only one method!

Microaggregation (good for risk, not so good for utility) followed by addition of noise (with variance $\Sigma_{\mathcal{O}} - \Sigma_{\mathcal{M}_{\text{microagg}}}$) has been shown to be effective in several settings

- Income data (Oganian & Karr, 2006)
- Variance estimation for Horvitz–Thompson estimators using replicate weights (Hang & Karr, submitted)
- Post-SDL editing (later)

The Risk–Utility Frontier

Approach to Selecting from Candidate Releases Choose from risk-utility frontier



Challenges (Cox, Karr & Kinney, 2011)

- One person's risk is another person's utility
- Extant utility measures tend to be either too broad—hence too blunt, or too narrow—tied to specific analyses

Using the WSSM to Evaluate Additive Noise

Previous thinking: Noise distribution should be the *data* distribution. This is good for utility, but bad for risk, and led to strategy of microaggregation followed by additive noise.

WSSM shows that Noise distribution should be the *measurement error* distribution

```
SELECTED PARAMETERS
Sample design:  SRS
WEB contact attempts:  1; CATI contact attempts:  2; CAPI contact attempts:  3
Numerical survey variable measurement error probability:  0.500
Categorical survey variable measurement error probability:  0.100
Numerical survey variable imputation method:  HotDeck
Categorical survey variable imputation method:  HotDeck
Numerical survey variable SDL method:  AdditiveNoise(0.15)
Categorical survey variable SDL method:  Swap(0.05)
Imputation of unit nonrespondents performed and reflected in H-T estimates
```

COUNTS

Population	Sample	WEB Resp	CATI Resp	CAPI Resp	Total Resp	Resp Rate
100000	5000	784	1448	1216	3448	0.690

The Results

```
KULLBACK-LIEBLER DIVERGENCES
Sample to population:  0.003289
Unit respondents to population:  0.004686
Final responses to population:  0.016306
Released data to population:  0.016870

HELLINGER DISTANCES
Sample to population:  0.048487
Unit respondents to population:  0.076446
Final responses to population:  0.228078
Released data to population:  0.219950

DISCLOSURE RISK
Pre-SDL: 2078.222619
Post-SDL: 5.083333
```

SDL in the Presence of Edit Constraints

Problem SDL can create violations of edit constraints

General Strategies

- Post-SDL editing
    - *Not-so-good ways:* Delete edit violators (problem: weights); Project violators onto feasible region (problem: points on boundary)
    - *Better way:* Replace violators using Kim, *et al.* (2013) "imputation subject to [linear] edit constraints" methodology
- Edit-preserving SDL
    - *Not-so-good way:* Alter the method so that it does not produce violators (problem: introduces bias [additive noise], infeasible in finite time [swapping])
    - *Better way:* partially synthetic data

Setting—1

Dataset 1991 Colombian Annual Manufacturing Survey Data

- 6521 records
- 7 variables: RVA (real value added), CAP (capital), SKL (skilled labor), USL (unskilled labor), RMU (raw material), SKW (skilled labor wages), USW (unskilled labor wages)

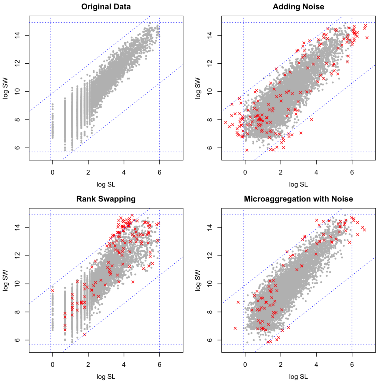
The Experiment

- Range and ratio constraints on all variables, derived from the data
- 7 SDL methods: additive noise; rank swapping; microaggregation via PC; microaggregation via z-score; microaggregation via PC followed by additive noise; microaggregation via z-score followed by additive noise; partially synthetic data

Setting—2

- Disclosure risk: linkage of masked data to original data, using composite variables
- Data utility
    - Kullback-Liebler divergence between original and masked data, assuming normality
    - Regression of log(RVA) on other variables
- Constraint-preserving imputation using Kim, *et al.* (submitted): if constraint is violated, all variables involved are imputed. Method is heavily Bayesian and computationally demanding, using a hit-and-run sampler.

SDL-Generated Edit Violations

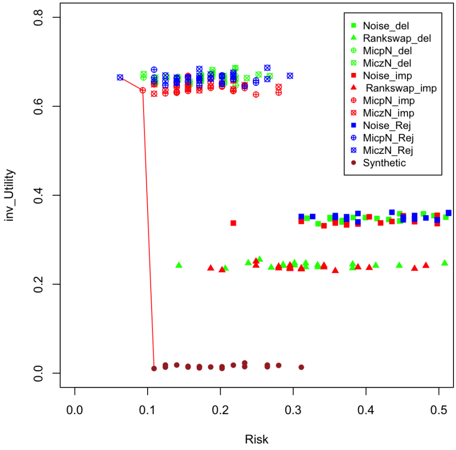


Masked variables are SL, USL and CAP

Which Methods Cause Problems?

SDL Method	% Violations
Noise	2.45
Rank Swapping	2.09
Micp	0.08
Micz	0.11
Micp + Noise	1.31
Micz + Noise	1.29
Partially synthetic data	0.0

Results


Noise_del
Rankswap_del
MicpN_del
MiczN_del
Noise_imp
Rankswap_imp
MicpN_imp
MiczN_imp
Noise_Rej
MicpN_Rej
MiczN_Rej
Synthetic
inv_Utility
0.0
0.2
0.4
0.6
0.8
Risk
0.0
0.1
0.2
0.3
0.4
0.5

Final Points

Where Next?

- Full, *scalable* integration of edit, imputation and SDL: need sound models for measurement error
- Mixed categorical and numerical variables
- Error localization

Acknowledgements

- Co-authors: Hang Kim (NISS/Duke), Jerome Reiter (Duke)
- NSF support: SES–1131897, "Triangle Census Research Network"